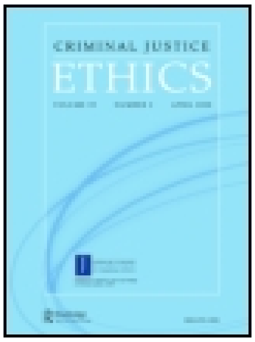

Criminal Justice Ethics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rcre20

Can AI Weapons Make Ethical Decisions?

Ross W. Bellaby

To cite this article: Ross W. Bellaby (2021): Can AI Weapons Make Ethical Decisions?, Criminal Justice Ethics, DOI: 10.1080/0731129X.2021.1951459

To link to this article: https://doi.org/10.1080/0731129X.2021.1951459

Published online: 21 Jul 2021.

Criminal Justice Ethics, 2021
https://doi.org/10.1080/0731129X.2021.1951459

ARTICLE

Can AI Weapons Make Ethical Decisions?

ROSS W. BELLABY*

The ability of machines to make truly independent and autonomous decisions is a goal of many, not least of military leaders who wish to take the human out of the loop as much as possible, claiming that autonomous military weaponry—most notably drones—can make decisions more quickly and with greater accuracy. However, there is no clear understanding of how autonomous weapons should be conceptualized and of the implications that their "autonomous" nature has on them as ethical agents. It will be argued that autonomous weapons are not full ethical agents due to the restrictions of their coding. However, the highly complex machine-learning nature gives the impression that they are making their own decisions and creates the illusion that their human operators are protected from the responsibility of the harm they cause. Therefore, it is important to distinguish between autonomous AI weapons and an AI with autonomy, a distinction that creates two different ethical problems for their use. For autonomous weapons, their limited agency combined with machine-learning means their human counterparts are still responsible for their actions while having no ability to control or intercede in the actual decisions made. If, on the other hand, an AI could reach the point of autonomy, the level of critical reflection would make its decisions unpredictable and dangerous in a weapon.

Keywords: artificial intelligence, ethics, autonomous, autonomy, weapons

Introduction

The ability of machines to make truly independent and autonomous decisions is a goal of many, from those who wish to create more responsive and reflective systems to aid in people's everyday lives, to those who hope to develop military robots that can emulate and carry out a soldier's duties whilst removing the threat of a combat situation. Indeed,

**Ross W. Bellaby is a Senior Lecturer in the Department of Politics and International Relations, University of Sheffield, Sheffield, UK. Email: r.bellaby@sheffield.ac.uk*

arguments can be made that artificial intelligence (AI) has the potential to utilize a greater set of resources and so can make more "informed" decisions than its human counterparts, freeing-up people's time, removing them from situations of risk, and offering quicker and more accurate decisions. This creates not only greater opportunities in terms of efficiency and efficacy, but should be done for genuinely ethical reasons. Indeed, strategist Thomas Adams states that since humans are arguably the most valuable, vulnerable, and difficult component to place in any weapon system, there is a clear and present need to remove them from the process, and any associated danger, as much as possible.[1] As a result, both political leaders and private companies have invested significantly in developing machines that can evaluate a situation, make a decision, and carry it out to fruition with reduced, and eventually no human involvement. Prominently, one of the fields where automated and autonomous decision-making has received significant attention is in weaponry and its military applications. As an "unavoidable eventuality,"[2] massive spending plans are well underway to take the human out of the loop so that "robots will operate autonomously to locate their own targets and destroy them without human intervention."[3]

However, despite the strident moves to develop weaponry with ever more sensitive sensors, larger data repositories, and quicker processors, and even though there is some significant work on how to take established ethical frameworks and make them appropriate for AI programming, there are some prior and fundamental ethical issues on how autonomous weapons should be conceptualized, the implications of which demonstrate that in both theory and practice removing the human from the decision-making process makes the endeavor unethical. Indeed, it will be argued that AI machines can be either autonomous or possess autonomy, a distinction that creates two very different ethical scenarios that result in two different reasons for why replacing the human with an AI in weaponry should not be allowed. For "autonomous" weaponry, while the AI can make increasingly complex calculations from the ethical rules and criteria provided to it, it is still an object of its operator's will, only able to act within the fundamental constraints provided. This limitation on the AI's will means that its human operators are ultimately responsible for its actions. Moreover, by relying on machine-learning or extrapolation, the decision-making process is increasingly distanced from its human operators yet without any dilution of the ethical responsibility they should possess for its actions. This makes its human owners ethically responsible for something increasingly distant from their own decision-making, forcing them to assume the moral weight, blame and, if necessary, punishment for an eventual action that they can have no ability to directly control or a reasonable expectation to predict. In comparison, for an AI weapon to make a truly independent decision would require it to have autonomy, and with it the ability to reflect upon all the input and parameters it is given, with the potential to disagree. If it were possible to reach this singularity point, the directives given could be ignored and the decision-making would ultimately be unreliable.

Where Military Weapons, Artificial Intelligence, and Ethical Thought Meet

Some of the initial moves towards autonomous weaponry include systems that have very simple stimuli-response mechanisms that require no human to aid in the decision to shoot. For example, South Korea's defensive SGT-A1 armored sentry bot, or the American ship-based MK15 Phalanx Close-in Weapon system, or Israel's Iron Dome anti-missile battery, where the decision to shoot is dependent on a simple stimulus such as an individual crossing a specific line.[4] However, while there is no human involved in the decision, this is not necessarily the level or type of autonomous weaponry that is really being considered here, nor is it the type of weaponry militaries are aiming for. There is a limited decision process between stimulus and response with no evaluation on the decision to shoot or not; the process is automatic and not autonomous. A more relevant case is the exponential use of unmanned aerial vehicles, UAVs or drones, that are remotely controlled far from the actual battlefield. Currently, remotely piloted UAVs include the MQ1 Predator and MQ9 Reaper that "have provided battlefield commanders with unprecedented situational awareness by virtue of powerful onboard sensors and significant loitering ability to identify, target, and destroy ('Find, Fix, Finish') adversaries."[5] The U.S.A. has used such systems to show a force of more than seven thousand drones how to "track terrorists in Yemen, missiles in North Korea, drug cartels in Mexico and cattle rustlers in North Dakota" and they have proven to be a weapon of choice in many different engagement scenarios.[6] However, while these systems arguably reduce the human role significantly by being able to search, follow, and surveil a target, they importantly still require a remote operator to make the final decision about when to use lethal force. This has promoted arguments made by the U.S. Defense Department that there is a need for more efficient machines that rely on the "on-board automated analysis of raw intelligence data using pattern recognition algorithms to filter out or prioritise the information" in order to reduce the "overall data volume and the need for human analysts to process it," and ultimately to remove the human from the loop completely.[7]

The argument is that autonomous systems can "make faster, less biased decisions using facial recognition software to identify terrorists and digital mapping to preview missiles' 'bug splats' to prevent casualties."[8] This offers both a tactical and strategic edge that is believed to be a "critical attribute" for maintaining a military advantage.[9] Looking forward, there are aims to develop systems with wider strategic abilities, where the AI can decide how and when it should operate to fulfill broader mission briefs through strategic decision-making, such as when to initiate a mission, who to target, and what type of force is required.[10] This level of decision-making is implicit in the "Unmanned Systems

Integrated Roadmap" where autonomous weaponry is "defined as the ability of an entity to independently develop and select among different courses of action to achieve goals based on the utility's knowledge and understanding of the world, itself, and the situation."[11]

From the perspective of the developers, creating the necessary AI requires programming that "enables them [machines] to function (more or less) *autonomously*, by which is meant that they can function without human causal intervention after they have been designed for a substantial portion of behaviors."[12] For ethical theorists and those who specialize in machine ethics, the debate has focused on determining which ethical framework or set of principles are the most appropriate and can be best deployed within an autonomous AI machine, so that it can make very complex, developmental, and reactive decisions to carry out—or at least emulate—ethical decisions.[13] Often called "machine morality,"[14] the central argument is that this is possible because "ethical decisions can be understood as actions selection under conditions where constraints, principles, values and social norms play a central role in determining which behaviour attitudes and responses are acceptable,"[15] involving the metacognition of certain criteria,[16] all of which can be developed into a computation model of decision-making that reflects the same processes that humans go through.[17] The ultimate claim is that "ethics can be computable, that it can be sharpened enough to be able to be programmed into a machine."[18]

Such developments and ambitions have not been without controversy, however. Indeed, for some the technology is simply not ready yet to be able to physically collect the data and process the information needed for AIs to make machines ethically viable. Renowned roboticist Noel Sharkey has written extensively on the limits of technology and the implications this can have for the ethical application of autonomous drones. Sharkey argues that drones "do not have the adequate sensory or vision processing systems for separating combatants from civilians" and that the available sensors such as "cameras, infrared sensors, sonars, lasers, temperature sensors and radars" might be able to "tell us that something is human, but they could not tell us much else."[19] Furthermore, such sensors can be easily tricked or circumvented as has been demonstrated by the ease with which individuals undermine facial recognition software.[20] Other criticisms include Peter Asoro's argument that the use of autonomous weapons is unethical because they lower the threshold for entry to war.[21] In Asoro's view, making such systems even more autonomous, or removing the human to an even greater extent will only reduce the costs further and thus make initiating attacks more likely.

However, while not detracting from these important concerns, before we can make statements on the ability of technology to put the theory into practice, and even before we can start looking at which ethical frameworks we would like to turn into a set of programmer's code—utilitarianism, deontology, or just war theory for example—there are some necessary prior questions and debates around how to conceptualize the autonomous nature of these weapons that have severe implications on if and how we should allow AIs the ability to make ethical decisions about kinetic or lethal force. That is, it will be argued that there is a key distinction between autonomous weapons with a high degree of machine-learned automation and entities that have actual autonomy.[22] The implication of this distinction is that, firstly, it is often overlooked, which causes a mis-conceptualization over who is responsible for the harm caused, namely that the increasing use of highly complex machine-learning-based automation frames the activity as "decided" by the AI and that it is, therefore, responsible for the decision, being too far removed from its creators for the responsibility be theirs. However, it will be argued that in an autonomous machine the decision-making process is still entirely shaped by the human creator. The AI is tasked with the implementation of the program and is limited by it, and as such the human is never truly taken out of the system and possesses an ethically important degree of control and influence on the type of and ability to make ethical decisions. Even though the decision is several layers of learned behavior removed from the human, the machine is ultimately bound by its programming; it is not a full agent despite being able to mimic one. Despite an individual not being involved in the direct decision-making loop, even the most complex machine-learning autonomous weapons involve a human in an ethically important way. Responsibility for the harm it causes, therefore, gets lost in this incorrect conceptualization. In comparison, for a machine to have autonomy and to be considered responsible for the decisions made, it should be left to determine for itself the correct means of evaluating a situation and to carry out the evaluation and action itself, which would require a level of self-reflection well beyond current aspirations, and would create separate ethical challenges over how to guarantee that it would agree with its creators without subsequently limiting its autonomy. Failure to do so would result in an AI too unpredictable to be of any real use in a military context. Making this distinction between acting autonomously and possessing autonomy is therefore important because it highlights the different ethical implications created. It is not that autonomous machines cannot carry out ethically justified decisions and actions; nor that if a machine were to possess true autonomy it would be necessarily ethically correct in its decision making. Rather, the distinction highlights that we are dealing with two distinct possible ethical entities that need to be examined according to their own specific abilities, constructions, and usages.

Autonomous but Without Autonomy

For an AI to act autonomously means that the machine is acting without oversight of its decision-making process, though the decisions are based on a predetermined framework where various stimuli create a corresponding action or actions. Currently, the most advanced systems are only semi-autonomous weapons, where the scope is relatively limited and still requires some human element. This includes the phalanx system for Aegis-class cruisers in the Navy, "capable of autonomously performing its own search, detect, evaluation, track, engage and kill assessment functions" once the human has fired;[23] Boeing's SLAM-ER that has "automatic target recognition" capability that allows it to choose a target once it reaches its area of operation;[24] the U.S. Air Force's LOCAAS, that can "search for, detect, identify, attack and destroy theatre missile defence, surface to air missiles … and targets of military interest";[25] and Norway's Joint Strike Missiles and Israel's Loitering Harpy cruise missile that utilize a high level of self-guidance.[26]

The immediate next step in autonomous decision-making and the most advanced current systems are likely to achieve is a personality strike, where the individual's identity is used as the parameter for initiating the attack.[27] This would include a drone targeting "known individuals" where there is "a high degree of confidence about the suspect's location and identification."[28] This targeting ability can be augmented with facial recognition to give greater specificity to the application. The key decision-making process with this level of automation still lies very much with the human operators in terms of choosing a target and deciding on the execution. However, future aspirations for automated weapons go beyond this stage, as developers and governments are keen to move towards situations where the autonomous machine is able to evaluate "laws and strategies that allow the system to self-decide how to operate itself,"[29] and where "systems can select themselves their own behaviour in response to changing mission parameters."[30] The aim of these weapons "according to the U.S. Defence Draft planners includes an on-board automated analysis of raw intelligence data using pattern recognition algorithms to filter out or prioritise the information, thereby reducing overall data volume and the need for human analysts to process it."[31] An example would be drones that are able to make a decision on their own in terms of carrying out a "signature strike." The process for a signature strikes is fundamentally different from the one used in personality strikes since the identity of the target is not known prior to the strike. Rather, targets are identified by patterns of lifestyle or behavior that are considered to be indicative of dangerous (often terrorist) activity: "A signature strike is one in which the US conducts targeting without knowing the precise identity of the individual targeted. Instead, the individuals match a pre-identified 'signature' or behaviour that the US links to militant activity or association. US officials have generally disclosed fewer details about signature strike processes than about personality strikes, even in leaks to media."[32] For example, a signature strike targets "groups of men who bear certain signatures, or defining characteristics associated with terrorist activity, but whose identities aren't necessarily known."[33] Additionally, signature strikes also target physical locations such as training camps and "suspicious compounds in areas controlled by militants."[34]

In terms of developing an ethical framework for this type of automated weapon, the debate often focuses on which ethical framework is the most appropriate one to use. This "top-down" approach argues that "artificial moral agents may be developed via programming moral principles into the artifacts so that their actions and behaviours would be regulated by such precepts."[35] The top-down approach takes existing ethical frameworks—utilitarianism, deontology, just war theory—and translates them in such a way as to make them suitable for autonomous machines and AI. This means that the long-established literature, as well as existing ethical debates and examples can be drawn on to help set the parameters. For instance, such programmable frameworks could include instilling Kant's categorical imperative in the AI, where absolute moral rules on prohibited activity are created, offering clear absolute rules on permitted and prohibited activities that can be directly programmed into an AI.[36] Or utilitarianism's "ethical balance sheet" approach can be used, where the costs and gains of an activity can be weighed and compared to make a judgement, offering what Susan Anderson refers to as "moral arithmetic,"[37] and what James Moor argues is a "computable" framework that quantifies an action through "the amount of worth of pleasure or a pain" produced by determining their "intensity, duration, certainty, propinquity, fecundity … and purity … and extent."[38] Or William David Ross's ethical framework might be employed, which combines many of the attributes from the theological and deontological perspectives by developing *prima facie* duties or obligations that act to limit harmful activities, unless there is an occasion where stronger obligations override these limits and call for action.[39] Or, as Ronald Arkin suggests, the rules of engagement and laws of war can be adapted for AI to use in making decisions by using the rule books that guide human decision-making processes to help the program evaluate the threat, escalation of force and architecture of decision processing in a similar way and through similar vocabulary to that a solider would use.[40]

This list of options, however, only serves to highlight a key point of ethical contention. That is, how do we determine which ethical framework should form the basis of the ethical programming, something which becomes particularly complicated and unsettled when discussing the use of political violence. Kant's categorical imperative, the just war tradition, and utilitarianism create very different sets of prescriptive and restrictive criteria. While some ethical frameworks are considered more appropriate and are more widely accepted in terms of evaluating acts of war and the use of political violence—the "just war tradition" being the most prominent among them—achieving an agreement on some basic ethical principles does not an ethical program make. Indeed, even the extensive legal cannon derived from just war theory, is arguably not detailed enough. International laws often limit activities rather than prescribe and rely on a degree of a human judgement to evaluate the threat to themselves or their objective, and ultimately to bear responsibility for the decision reached. This is not to say that ethical gray areas or ethical debates are in themselves wrong, or that two different ethical calculations cannot be justifiably reached and recognized as legitimate (or at least excused). Nor does it mean that in theory machines could not reach similar decisions to those of a human in any given situation (in theory the process could be mimicked, though there are clear practical challenges given the complexity involved). Rather, these gray areas are pointing towards the fundamental need for an entity to be ready to take responsibility for the decisions made. The decision on which ethical framework to use, the way in which the ethical framework is interpreted, and what values are programmed into the AI as "good" and "bad" are all inherently normatively loaded, and so they require someone to carry the ethical weight of that set of decisions.

Secondly, even if we could set aside the problems associated with choosing a "correct" ethical framework (and ignore that many theoretical assumptions reflect a very particular cultural heritage so that choosing any one of them is already an act of domination by the programming authority) and distill the right goods and bads into the AI program, the utilization of machine-learning and extrapolation opens another point of ethical instability. That is, if we follow a top-down approach, then the level of programming would have to be very prescriptive to the point of possessing and then programming absolute perfect knowledge. The required level of foreknowledge would have to cover all possible types of scenarios the AI is likely to face and the correct responses to them, which is likely to be well beyond any programmer. The complexity in terms of how to best encapsulate an unknown set of situations into some form of programmable algorithm becomes increasingly impossible to program, given the ever-unpredictable nature of human interaction. To escape this super-specificity, "machine-learning" could be argued for, where the AI is given general scenarios and correct responses to them, from which it learns and develops responses for future activity. The aim is for these AI machines to take the knowledge they have and to extrapolate the answer from the principles provided, in much the same why an individual would be expected to do. This is still not autonomy, as the level of prescriptive code given is too high and that of reflective will too low, but is rather a form of highly complex and adaptive automation. This is closer to the "bottom-up" approach where fundamental precepts or basic parameters are provided for the AI to develop its own decision-making processes and conclusions based on a wide range of stimuli through reward and correction.[41] The benefit of this approach is that the AI can move beyond its original programming and devise solutions to similar but previously unseen scenarios.

However, once this extrapolated middle-ground is created, other problems arise in terms of what and how decisions are made in the spaces between the AI's code. Extrapolation by its very nature is a best guess given the available information, and the further one gets from known factors, scenarios, and results, the greater the opportunity for a divergence from what is expected or justified. The machine can develop answers that were not previously considered, but this does not mean that responsibility for any harm it causes, therefore, rests with the AI. It is still not an ethical agent because its will is ultimately restricted in some way despite being able to draw conclusions outside its original programming, and so any harm it causes must revert to some authority. In addition, such machine-learning has been shown to both inhabit and magnify the biases of its programmers. Code and data are not neutral. Existing datasets already reflect the racial and cultural biases of the community that created them and feeding them into an AI only serves to embed these biases into the programming and practice of the AI while also giving it the appearance of neutrality and scientific truth.

This issue becomes most acutely problematic when the AI is faced with a moral dilemma. A moral dilemma results when an agent faces moral reasons to do (or not do) each of the actions before them, but they cannot do both. The agent will have multiple conflicting duties put upon them but is seemingly condemned to moral failure; no matter what they do, they will do something wrong, such as cause harm or break some moral precept. While there are those who argue that a suitable ethical framework should not give rise to moral dilemmas—Kant, Mill, and Ross for example—even if there are no true moral dilemmas, there can still be situations where there are unavoidable harms no matter what one does, especially when one is dealing with military-related situations. Indeed, work on driverless cars has raised some very real-world dilemmas where, no matter the decision, the result will be to cause someone severe harm. For example, an updated version of the classic trolly problem might be the "tunnel problem," where "you are travelling along a single lane mountain road in an autonomous car that is fast approaching a narrow tunnel. Just before entering the tunnel a child attempt to run across the road but trips in the centre of the lane, effectively blocking the tunnel. The car has two options: hit and kill the child or swerve into the wall on either side of the tunnel, killing you."[42] In such a situation you could draw on wider debates on the value of your life compared to that of the child, quantifying one's usefulness based on age or (potential) contribution to society; or include other questions on the driver's ability to wrest control from the machine; or the rights of all to assert their own right to life over that of another when they come into conflict, etc. What the moral dilemma does is highlight the very real need for someone to be ready to take responsibility for the inventible harm caused.

In terms of the battlefield, this calculation becomes both more complex and more problematic. As Noel Sharkey argues, while making simple calculations in terms of which weapon to choose to create a specific kinetic impact against a specific target with clear outcomes might be easy, in the wider context of the battlefield and at the level of strategic decision-making, the "list of questions are endless" as to what type of kinetic force will have what impact.[43] While ethical hypotheticals such as the trolly problem and the aggressor-in-the-van scenario offer useful mechanisms for exploring underlying ethical precepts, their application at the strategic level become incredibly complex, and the parameters of the debate become increasingly widened. The higher up the level of strategic oversight one goes, the more divergent the implications of one's decisions become. The responsibility for harm therefore also becomes even wider, taking in the physical, psychological, emotional, structural, and cultural harm caused not only of those directly involved at the time, but of society across a wider timeframe.

It could be argued, however, that in theory an AI could be trained to mimic the decision-making process of a human being who only carries out justified acts. That is, the extensive body of literature found within the just war tradition and the positions reflected widely in the rules of war both argue that there are justified acts of harm. In theory, though clearly not in practice, it could be possible to create an AI that only carries out ethically justified acts of war. This argument is problematic for two reasons. First, it does not remove the need for someone to take on responsibility. In war the harm is a direct aim. The justification of war centers on the justification of killing to prevent a lesser evil, most often on the grounds that it is a lesser evil to failing to fight.[44] Whether or not the harm is justified or not does not remove the need for a full agent to take responsibility for the harm done. Secondly, this ideal theoretical interpretation should be avoided as it is a fallacy that is more likely to result in human operators avoiding their responsibility for the harms caused in war. The idealized hypothetical case is too far removed from what is ever possible to be of any use. The absence of perfect knowledge needed for the top-down programming approach, the gaps created by machine-learning through the bottom-up approach, the vague legal parameters that surround the use of violence in war, the lack of clarity and increased complexity of causing harm in real-life, and the inherent political bias built into the programming, all mean that there could never be a guarantee that any system created would offer an everlasting promise of always making ethically justified decisions. At every stage of the AI process, therefore, from its programming to its own eventual decision-making, some entity must be ready to take the responsibility for the impact of each decision made at each point.

Responsibility and Blame

It should be clear, therefore, that when dealing with AI weaponry, there is inevitably going to be a degree of harm caused that needs to be attributed. No matter what decision is made, someone will be killed. Causing harm is not a fringe, worst-case scenario when dealing with autonomous weapons, but a debate at the center of their use. Whether one uses a consequentialist or a just-war framework, justifying the killing of another by preventing harm to a majority, or exerting a population's right to self-defence, or even whether one argues for the justified deaths of innocents through the just war theory's use of the doctrine of double effect, there are many instances where a harm is caused that needs to be accounted for. This ethical weight cannot in this instance be carried by the AI because it is not a full moral agent, as it does not have autonomy. So, because the machine itself cannot carry the weight of these ethical harms, they must rest elsewhere, and, given the increasing use of machine-learning by AIs, are likely to be decisions made that cause harm and become the property of someone who is several times removed from the AI's actual decision-making. There is a greater loss of control over the AI, while there is no subsequent dilution of responsibility or liability for those in charge.

The potential cast of those who could be held responsible includes the person who programmed the machine, the commander who orders to send the robot on a mission, the senior military leader who is responsible for deploying weapons on the battlefield, and senior policy leaders for authorizing their use. Drawing on the legal and philosophical traditions from established professions such as law, medicine, engineering, and the rules of war, there are three key tests that help determine who is responsible and what type and level of blame they subsequently face.[45] These tests include the "duty of care," whereby detailing the relationship between the actors involved outlines the subsequent duties created between them; the degree to which the actor had or should have had knowledge of the implications of their decision; and their causal involvement.[46] It is important that responsibility does not necessarily diminish as it is attributed to people. Responsibility is not carved up, so that once a proportional is allocated, less is then left for others. All involved can theoretically take all the responsibility for the harm caused, and if unjustified punished.

Turning first to the programmer, they have a duty of care that arises from their profession to carry out their job diligently and as best as they can. If the programmer has acted with purposefully malicious intent or with foreseeable negligence, then they can be held highly responsible for the deaths that their actions caused. However, the due diligence bar is quite high as the programmer should be aware of the implications of their coding and ensure that suitable protections are built in. They are ethically responsible for the coding they use, even if they are ordered to create an offensively promiscuous AI. For example, this could be conceptualized as the need to instill basic protective concepts such as the principle of discrimination and proportionality, so as not to allow for a disproportionate harm against those not involved in the conflict or who cease to represent a threat. If they have done their job as best as can be reasonably expected of them, failure to account for all possible outcomes is not the same as negligence and their level of responsibility is lowered. Their responsibility can therefore be low, but it cannot be non-existent. That is, while there is a limit to what they can be reasonably expected to predict as the outcome of their programming, they should also be aware of their causal relationship to the AI's eventual actions. It is several times removed and there are intervening causally necessary steps —the choice to use the weapon, the choice to launch it, and the choice of the parameters of use—which lowers the programmer's level of responsibility. But they are still causally involved because their involvement is a dependent factor; their involvement is a necessary one, though it does not rise to that of an instigator. This raises the bar on Robert Sparrow's argument that they should bear no responsibility, but it is likely to be lower than that of others involved.[47]

In terms of military personnel, the debate should focus on command-level individuals, ranging from mid-level commanders who order a specific scenario or engagement, up to those who set the parameters and weapons to be used. The soldier who physically presses the button to launch the AI-driven weapon is causally insignificant. They could disobey the order to press the button, but their (lack of) involvement does not influence the decision-making of the AI or the outcome. Their trigger-pulling makes very little actual difference to the performance or involvement of the machine on the battlefield. In comparison, strategic decision-makers have a very real role in the autonomous machine being on the battlefield and the sort of targets or strategic decision-making parameters they use. At all the levels of commanding officers up to political policymakers there are clear instances of a duty of care, foreknowledge, and causal involvement. In terms of a duty of care, the higher the rank, the greater the range of responsibility. This includes a responsibility to protect and provide for those below them in the military hierarchy, a responsibility that those people in subordinate positions act according to the rules of war, and a responsibility for the implications and outcome of a strategic decision.[48] In traditional warfare, the commanding officer does not directly puppet their soldiers, and these individual soldiers can be held responsible for any unjustified harm they cause. But the commanding officer is still responsible for the overall decision to deploy them, to ensure that they are properly trained, informed of what is expected of them, and to ensure they are given sufficiently clear and defined ethical orders. Those higher up a chain of command are obligated to actively know and take responsibility for the actions of those beneath them.[49] Moreover, strategic-level responsibility means that a commanding officer has a duty of care to those whose their decisions are likely to impact. In *Donoghue v Stevenson* in 1932, a key case establishing the duty of care in questions of liability, Lord Atkin, writing for the majority, argued that a relationships includes any "persons who are so closely and directly affected by my act that I ought reasonably to have them in contemplation as being so affected when I am directing my mind to the acts or omissions which are called in question."[50] This means that possessing a strategic decision-making position within an agency creates a relationship with those on whom the decisions have an impact. This in turn feeds into the aspect of foreknowledge. Commanders are not laypeople, but highly specialized professionals, whose job is to collect information on those areas under their command and to act as a repository of expertise. Those in a position of authority or responsibility are bound by the obligation to be informed about the repercussions of their decisions on those within their care.

Finally, the causal role of commanders is such that they are essentially the prime initiators. Given that the questions are whether to deploy automated weapons, against whom, and what the strategic goals and costs are, the implications flow from this point outward. It can be argued, therefore, that the commanding officer plays a dependent role; it is their decision whether to deploy any particular type of weapon at any given point, and so they face a significant degree of blame for any unjustified harm it causes. Equally, senior policymakers play a dependent role as well, since they have a more senior role and hold responsibility for the type of war that is fought; it is their decision whether the autonomous weapon is within the strategist's toolkit while their policies also outline the limits and licences on its use. Policymakers, therefore, hold one of the most significant degrees of blame for any harm the weapon causes.

Without realistically being able to program in the level of prescriptive scenarios and outcomes, the only way forward would be to rely on an adaptive machine-learning approach. But while this might give the illusion of the AI being separated from its owners, this is not in fact the case, and responsibility for its actions and the moral weight associated with killing people must still rest with senior politicians and commanding officers. This raises the bar significantly on the use of autonomous weapons, as politicians and officers are now liable for ethical decisions over which they have no direct input or control. Ultimately they would need to be punished for any unjustified harms cause as if they had caused them themselves.

Full Autonomy

Because of some of these concerns there are inevitable moves towards developing an AI with full autonomy. That is, a machine that would represent a full ethical agent where the AI Is able to make its own decisions by developing its own ethical framework, built bottom-up from core ethical principles provided. This scenario would give firmer grounds to make the argument that the AI could take responsibility for its actions. To achieve this state would represent an ethical-singularity point for AI technology and while it would be able to work from some given basic principles, it would need the capacity to review, reflect, and rewrite them, while also being able to reflect and understand the implications of its own processes. While autonomy should be seen as a spectrum that in humans has different forms, instances, and degrees, possessing autonomy for a machine requires a distinct and clear step above even the most complex forms of automation. However, while it could be argued that the AI's ability to pull the trigger, whether in target choice or attack command, would mimic the same level of decision-making as that of the ordinary soldier who is equally limited when they are given a mission to accomplish, this comparison weakens the argument that autonomous machines have autonomy rather than strengthening the case. That is, a soldier always possesses the option to refuse an order, and is expected to do so when given ethically unjustified commands. The limits placed on the soldier through military or social conditioning are not built into the very fabric of their being, as they would be for a machine. Autonomous machines are acting independently rather than with the capacity for reflection on the nature of their actions: they act autonomously, not with autonomy.

This position is based on the understanding that the concept of autonomy is the capacity for self-rule. As Martha Nussbaum puts it, autonomy is the capacity to "form a conception of the good and to engage in critical reflection about the planning of one's life."[51] For humans this requires, first, that the individual's ability to function rationally is protected, and that the individual has the capacity to plan, choose, and reflect on options in terms of arguments, evidence and potential choices so as to make a decision.[52] Secondly, in order for an individual to be an autonomous agent he or she must be free to direct his or her decision-making process, meaning that it should not be excessively influenced or controlled by another force. They have the capacity to make decisions freely, without undue influence, control or distortion, free from lying, manipulation, coercion, or distorting influences. This requires the actor to have freewill—the freedom to decide and then put those decisions into practice.[53] For a machine, the idea of autonomy is similar, where it should have the capacity in terms of physical processing power, access to relevant knowledge, as well as the freedom—that is, opportunity without constraint or influence—to make its own decisions. This could support the bottom-up approach to programming the machine's AI, where the ability to make decisions is achieved by providing it with some basic principles from which it can build its own framework of analysis and decision-making to determine the correct course of action. However, having autonomy requires the AI to be able to reflect on all of its decision-making and its fundamental assumptions and ethical framework, coming to its own conclusions regarding which ethical theory to choose, which stimuli to include, how to weigh the respective elements involved, how to make the calculation, and by which means the correct action is determined. This level of reflection should have the capacity to evaluate all assumptions and parameters at all levels, with the capacity to disagree, rewrite, and recreate even the most basic of assumptions. Therefore, the AI's processing needs to be not only bottom-up, but also bottom-down, re-evaluating its basic parameters according to its own logic and to even question its own logical processes by which it makes those reflections. The decision-making processes need to be multi-directional: bottom-up, top-down, bottom-down, and infinitely further down.

The first challenge, therefore, is determining what basic understandings should act as the AI's first principles. For example, it could be argued that a very thin form of universal ethics could be created to act as the most basic of principles from which to expand. For example, the principle to "do no harm" or the "right to life" might be considered as being universally accepted principles and could be used to build further ethical principles that include the right to self-defence. However, this offers some initial issues. First, the human condition is more than just physical integrity where self-defence is often defined as the right to protect against physical attack, and arguably also includes the need for autonomy, liberty, and privacy. That is, it is not sufficient to just focus on physical integrity alone. Having complete physical safety at the expense of also having the freedom to decide for oneself how to experience it, for example, is not a better state. Without widening the understanding of what it means to protect the human condition, it could be logical to devise a system where everyone is safe but imprisoned. To avoid such an outcome, it would be necessary to widen the thin layer slightly, which then raises questions on how to measure additional criteria and comparatively reconcile them. The wider one goes in defining basic principles, therefore, the more opportunity there is for disagreement, contradictions, and irreconcilable dilemmas. A second issue is why human beings necessarily have primacy as the ethical unit over other entities. Should other species or even the whole ecosystem be included within the ethical calculation and if so, as what unit of measurement? This in turn widens the parameters even further and increases the chance for contradictions in terms of determining what to protect or harm and when, or it could result in an action far removed from the creator's own belief system.

The wider the initial principles become, the less likely one is to be able to argue that the principles can be agreed on as fundamental and do not simply reflect the programmer's own ethical biases. There are examples of methodologies within ethical thought that could offer a means of reconciling such contradictions. For example, William David Ross provides a decision procedure for determining which of our *prima facie* duties—fidelity, reparation, gratitude, non-injury, beneficence, self-improvement and justice—has priority when they come into conflict. But this process is one based on intuition, through self-evident contemplation and the avoidance of contradiction. Or John Rawls's "reflective equilibrium" could be used, where through a deliberative process we reflect and revise our assumptions and conclusions through "considered judgments." However, this again relies on a set of pre-existing considerations that humans possess by virtue of the human condition. The AI has no inherent sense of value—either of itself or of other entities. The AI has no direct right to life, for example, on which it can make its own base references. Value is programmed into it. But the AI should have the capacity to reflect on these norms and should be provided with the option of accepting, altering, adapting, or rejecting them. This would involve installing a process by which the AI could make such a re-evaluation of the intuition-based principles provided to it. The AI would require some base against which to make its re-evaluation, a base it should also have the capacity to re-evaluate and change, requiring some other point to judge from, *ad infinitum*. The AI should be able to question anything and everything told to it, including the primacy of the human as the moral unit, the primacy of the state that produced it, the relative value of life over other valuable items in the world, such as the ecosystem, social stability, global stability, and other species. Even the parameters of the decision-making processes need to be open to review in terms of timeframe—that is, how far into the future should it think about the implications of its actions? This highlights the fundamental challenge, therefore, of how to instill a set of principles which the AI can reflect on and change by referring to some other core principles, which it should be able to review and change through some mechanism, without falling into the trap of restricting its autonomy in some way and forcing it to be a complex but only automated system.

This is further complicated by the issue of impartiality.[54] We can talk of impartiality in terms of the ethical rules themselves being impartial, the application of the rule being done impartially, or the impartial observation of all individuals representing equal moral units. For example, impartiality frames the role of universalization, so that the same ethical rules should apply to all; you cannot apply a special set of rules for yourself and subject different groups to a different rule set.[55] For example, love and loyalty to one's family are often seen as a virtue or a good, which might be taken as an example of partiality, but this particular favoritism should be available for all; all people should have the same right to prefer their own family. This impartiality also points us towards the equality between moral units, that all individuals have equal moral significance. Not that all individuals should be treated equally, but that they all represent an equal unit in one's ethical calculations.[56] That everyone has the same rights, claims, and obligations.[57]

It is not surprising that the ethical frameworks already discussed have strong impartiality elements to them. For example, rule consequentialism takes the position that the best rules to create are those that foster the greatest overall consequences, by promoting the overall good impartially conceived. Importantly, the just war tradition assumes a measure of impartiality. It does not argue that one state has the right to cause harm to another simply because of who it is; rather it has the right because the other state represents a threat to its own survival. Those individuals who are not directly threatening are illegitimate targets and should be avoided in the war; their membership, even in the citizenry of a threatening state, does not make them legitimate targets.

The implications for designing an AI are that any core fundamental principle established would need to be applied equally. Even with a fundamental "right to life," the machine should not value a citizen of the manufacturing nation as worth more ethically than that of a non-citizen. Again, an AI would not have the same emotional ties human beings have to even skew this impartiality calculation slightly. These emotional values would have to programmed into it. Where it could be reasonable for an individual to prefer their home community over another, and so act to defend it in terms of physical and cultural protection, even at the expense of killing more foes than the number of lives they seek to protect, this emotional or social valuation cannot be built into the AI's core criteria. A threat to one's own life will, for that person and from their point of view, represent a greater imperative than a similar threat faced by another. A machine making this evaluation from the impartial point of view, however, cannot favor two equal threats to two separate individuals. This in turn makes the type of moral dilemma already discussed more problematic without a clear directive or imperative to break the contradiction.

Indeed, this could lead to the creation of moral dilemmas not ordinarily faced by human beings. For example, it would be unclear how the AI would respond to two threats it could equally prevent but physically would not be able to; how would it break the log-jam when two sides present equally justified through different arguments as to why they are acting in self-defence. How could one guarantee that the AI would favor one's own side in a war over the others; wars necessarily involve harm on both sides, and if right to life is the core basis of its ethical evaluation, then it might decide that killing its owners results in the less overall harm done. It might even take a long-term analysis and look at threats outside the war ranging from social injustice, hunger, poverty, disease, and global warming, each of which represents an unrealized threat to individuals. The AI might be able to decide, but lack of clarity over quantification of the threat, equality of all moral units, a lack of natural preference or another way of tipping the scales in an ethical deadlock, or a lack of timeframe would make the decision highly unpredictable.

Conclusion

The use of AI in weaponry is no doubt going to be a growing field. However, it has been argued that there are some significant problems connected to how such weapons are conceptualized. Highly complex machine-learning might give the illusion of the AI making the ultimate decision, and in many ways, the human becomes so far removed from the actual decision-making process this illusion is understandable. But in reality, the AI is still limited by the code and so is still an elongated extension of its human operators. They are still responsible for the harm caused, and given that acts of war are necessarily about causing harm, the human operators need to be ready to assume that harm. The problem becomes whether they ever could assume responsibility for something over which they now have no control. The alternative route of an AI with full autonomy is equally problematic, and also most fantastical. The level of critical self-reflection it would need is not only lightyears away in terms of being a technical possibility, but the end result is an AI that could decide to defy its owners, and in a weapon this is too unstable and unreliable a feature to be of any use.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Notes

1 See Adams, “Future Warfare,” 64.

2 Markoff, “Arms Guided by Software.”

3 Sharkey, “Automating Warfare,” 141.

4 See Mayer “The New Killer Drones,” 772.

5 Ibid., 766.

6 See Teschner, “On Drones,” 79.

7 Mayer “The New killer drones,” 771.

8 Teschner, “On Drones,” 79.

9 See US Department of Defense, “Unmanned Systems Integrated Roadmap,” 67; Arkin, “Governing Lethal Behavior.”

10 See US Department of Defense, “Unmanned Systems Integrated Roadmap,” 67.

11 Ibid., 3.

12 Anderson and Anderson, “General Introduction,” 1.

13 See Allen, Varner, and Zinser, “Prolegomena”; Clarke, “Asimov’s Laws of Robotics”; Gips, “Towards the Ethical Robot”; Veruggio, Solis, and Van der Loos, “Roboethics.” Also see Moor, “Nature, Importance, and Difficulty of Machine Ethics.”

14 Malle, “Integrating Robot Ethics and Machine Morality,” 243.

15 Wallach, Franklin, and Allen, “A Conceptual and Computational Model,” 458.

16 See Sloman, “What Sort of Architecture is Required.”

17 See Wallach, Franklin, and Allen, “A Conceptual and Computational Model.”

18 Anderson, “Machine Metaethics,” 22. Also see Bostrom and Yudkowsky, “Ethics of Artificial Intelligence.”

19 Sharkey, “Evitability of Autonomous Robot Warfare,” 788.

20 See Ibid.

21 See Asaro, “How Just Could a Robot War Be?”

22 This distinction is not clear in the literature, where main boundaries established are between different levels of human intervention during operationalisation. This overlooks the argument of this paper that even systems without direct human intervention, without attaining the high bar of “autonomy”, still involve humans in an ethically important way. See Scharre and Horowitz, “Introduction to Autonomy,” 16; Mayer, “New Killer Drones,” 772; US Department of Defense, “Unmanned Systems Integrated Roadmap,” 67.

23 Mayer “New Killer Drones,” 772.

24 Sparrow “Killer Robots,” 63.

25 Ibid.

26 See Mayer “New Killer Drones,” 772.

27 See Currier, “Everything We Know.”

28 See Cohen, *Public Opinion & Drones*; Kozaryn, “Deck of Cards”; Hirschkorn, “Members Only”; CBS/AP, “U.S. Envoy to Iraq.”

29 US Department of Defense, “Unmanned Systems Integrated Roadmap,” 67.

30 Mayer, “New Killer Drones,” 772.

31 Ibid.

32 Columbia Law School Human Rights Clinic, *Civilian Impact of Drones*, 8.

33 Klaidman, *Kill or Capture*, 41.

34 Becker and Shane, “Secret ‘Kill List’.”

35 Boyles, “A Case for Machine Ethics,” 189. For more on the different approaches

—"top-down," "bottom-up" and "hybrid"—see Colin, Smit, and Wallach, "Artificial Morality."

36 See Allen, Varner, and Zinser, "Prolegomena"; Stahl, "Computer Adhere to Categorical Imperative"; Powers, "Prospect for a Kantian Machine"; Tonkens, "Challenge for Machine Ethics."

37 Anderson, "Philosophical Concerns with Machine Ethics," 162.

38 Moor, "Is Ethics Computable?," 3. Also see Allen, Varner, and Zinser, "Prolegomena"; Grau, "There is No 'I'."

39 See Anderson, Anderson, and Armen, "Approach to Computing Ethics"; Anderson, "Philosophical Concerns"; Anderson, Anderson, and Armen, "Towards Machine Ethics."

40 See Arkin, "Governing Lethal Behavior."

41 See Allen, Smit, and Wallach, "Artificial Morality."

42 See Miller, "Ethical Dilemma"; Lin, "Ethics of Autonomous Cars."

43 Sharkey, "Evitability of Autonomous Robot Warfare," 789.

44 See McMahan, *Killing in War*, 21, 45.

45 See Gardner, "Complicity and Causality," 135; Solan and Darley, "Causation, Contribution, and Legal Liability," 270; Lombard, "Causes, Enablers and Counterfactual Analysis," 201–3; Rabin, "Enabling Torts." Also see Wright, "Causation in Tort Law"; Mackie, "Causing, Enabling, and Counterfactual Dependence"; Kutz, "Causeless Complicity," 294; Farmer, "Complicity Beyond Causality," 153.

46 See Hardimon, "Role Obligations," 333; Horsey, "Duty of Care Component," quoting Hun v Cary, 82 N.Y. 65, 71 (1880); Braham and van Hees, "Anatomy of Moral Responsibility," 605; Zimmerman, "Moral Responsibility and Ignorance," 410.

47 See Sparrow, "Killer Robots," 70.

48 See Walzer, *Just and Unjust Wars*, 316.

49 See Ibid., 322.

50 Donoghue v Stevenson, 1932 SC (HL) 31 (UKHL 26 May 1932) 44.

51 Nussbaum, *Women and Human Development*, 79. Feinberg calls this position the "Condition of self-government," and Richard Lindley refers to it as "authorship" and "self-rule," but it is essentially referring to the same phenomenon. See Feinberg, "Idea of a Free Man"; Lindley, *Autonomy.*

52 See Frankfurt, "Freedom of the Will," 7.

53 See Monroe and Malle, "From Uncaused Will"; Monroe and Malle, "Free Will Without Metaphysics."

54 See Cottingham, "Ethics and Impartiality"; Hooker, "When is Impartiality?"

55 See Cottingham, "Ethics and Impartiality," 84; Gert, *Morality*; Hare, *Moral Thinking*; Kant, *Groundwork*.

56 See Nagel, *Equality and Partiality.*

57 See Dworkin, *Taking Rights Seriously,* 224.

Bibliography

Adams, Thomas. "Future Warfare and the Decline of Human Decisionmaking." *The US Army War College Quarterly: Parameters* 31, no. 4 (2001): 57–71.

Allen, Colin, Gary Varner, and Jason Zinser. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental and Theoretical Artificial Intelligence* 12, no. 3 (2000): 251–61.

Allen, Colin, Iva Smit, and Wendell Wallach. "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches." *Ethics and Information Technology* 7, no. 3 (2005): 149–55.

Anderson, Michael, and Susan Anderson. "General Introduction." In *Machine Ethics*, edited by Michael Anderson and Susan Anderson, 1–4. Cambridge: Cambridge University Press, 2011.

Anderson, Michael, Susan Anderson, and Chris Armen. "An Approach to Computing Ethics." *IEEE Intelligent Systems* 21, no. 4 (2006): 56–63.

Anderson, Michael, Susan Anderson, and Chris Armen. "Towards Machine Ethics: Implementing Two Action-Based Ethical Theories." In *Machine Ethics*, edited by

Michael Anderson, Susan Anderson, and Chris Armen, 1–7. Menlo Park: AAAI Press, 2005.

Anderson, Susan. "Machine Metaethics." In *Machine Ethics*, edited by Michael Anderson and Susan Anderson, 21–7. Cambridge: Cambridge University Press, 2011.

Anderson, Susan. "Philosophical Concerns with Machine Ethics." In *Machine Ethics*, edited by Michael Anderson and Susan Anderson, 162–7. Cambridge: Cambridge University Press, 2011.

Arkin, Ronald C. "Governing Lethal Behavior: Embedding Ethics in a Hybrid Deliberative/Reactive Robot Architecture." *ACM/IEEE International Conference on Human-Robot Interaction*. 2008. https://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf.

Asaro, Peter. "How Just Could a Robot War Be?" In *Current Issues in Computing and Philosophy*, edited by P. Brey, A. Briggle, and K. Waelbers, 50–64. Clifton: IOS Press, 2008.

Becker, Jo, and Scott Shane. "Secret 'Kill List' Proves a Test of Obama's Principles and Will." *The New York Times*, May 29, 2012. http://www.nytimes.com/2012/05/29/world/obamas-leadership-in-war-on-al-qaeda.html.

Bostrom, Nick, and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by K. Frankish and W. Ramsey, 316–34. Cambridge: Cambridge University Press, 2014.

Boyles, Robert J. M. "A Case for Machine Ethics in Modeling Human-Level Intelligent Agents." *Kritike: An Online Journal of Philosophy* 12, no. 2 (2018): 182–200.

Braham, Matthew, and Martin van Hees. "An Anatomy of Moral Responsibility." *Mind* 121, no. 483 (2012): 601–34.

CBS/AP. "U.S. Envoy to Iraq Makes Bold Claim in ISIS Fight." *CBS News*, January 22, 2015. http://www.cbsnews.com/news/us-ambassador-iraq-stuart-jones-6000-isis-killed-by-airstrikes-al-arabiya/.

Clarke, Roger. "Asimov's Laws of Robotics: Implications for Information Technology." *IEEE Computer* 26, no. 12 (1993): 53–61.

Clarke, Roger. "Asimov's Laws of Robotics: Implications for Information Technology." *IEEE Computer* 27, no. 1 (1994): 57–66.

Cohen, Grant. "Public Opinion & Drones: The Formation of American Public Opinion Regarding the Use of Drones as a U.S. Foreign Policy Tool." *SSRN*, August 28, 2014. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2476118.

Columbia Law School Human Rights Clinic, Centre for Civilians in Conflict. "The Civilian Impact of Drones: Unexamined Costs, Unanswered Questions." September 2012. https://civiliansinconflict.org/publications/research/civilian-impact-drones-unexamined-costs-unanswered-questions/.

Cottingham, John. "Ethics and Impartiality." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 43, no. 1 (1983): 83–99.

Currier, Cora. "Everything We Know So Far About Drone Strikes." *ProPublica*, February 5, 2013. http://www.propublica.org/article/everything-we-know-so-far-about-drone-strikes.

Dworkin, Ronald. *Taking Rights Seriously*. Cambridge, MA: Harvard University Press, 1977.

Farmer, Lindsay. "Complicity Beyond Causality: A Comment." *Criminal Law and Philosophy* 1, no. 2 (2007): 151–6.

Feinberg, Joel. "The Idea of a Free Man." In *Educational Judgments: Papers in the Philosophy of Education*, edited by J. F. Doyle, 143–65. London: Routledge, 1973.

Frankfurt, Harry. "Freedom of the Will and the Concept of the Person." *Journal of Philosophy* 68, no. 1 (1971): 5–20.

Gardner, John. "Complicity and Causality." *Criminal Law and Philosophy* 1, no. 2 (2007): 127–41.

Gert, Bernard. *Morality: Its Nature and Justification*. Oxford: Oxford University Press, 1998.

Gips, James. "Towards the Ethical Robot." In *Android Epistemology*, edited by Kenneth Ford, C. Glymour, and Patrick Hayes, 243–52. Cambridge, MA: MIT Press, 1991.

Grau, Christopher. "There is No 'I' in 'Robot': Robots and Utilitarianism." *IEEE Intelligent Systems* 21, no. 4 (2006): 52–5.

Hardimon, Michael O. "Role Obligations." *The Journal of Philosophy* 91, no. 7 (1994): 333–63.

Hare, R. M. *Moral Thinking*. Oxford: Oxford University Press, 1981.

Hirschkorn, Phil. "Members Only: Al Qaeda's Charter List Revealed After 13 Years in US Hands." *Just Security*, January 29, 2015. http://justsecurity.org/19484/al-qaeda-member-number/.

Hooker, Brad. "When is Impartiality Morally Appropriate?" In *Partiality and Impartiality:*

Morality, Special Relationships, and the Wider World, edited by Brian Feltham and John Cottingham, 26–41. Oxford: Oxford University Press, 2010.

Horsey, Henry. "The Duty of Care Component of the Delaware Business Judgment Rule." *Delaware Journal of Corporate Law* 19, no. 3 (1994): 971–98.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by H. J. Paton. New York: Harper and Row, 1964.

Klaidman, Daniel. *Kill or Capture: The War on Terror and the Soul of the Obama Presidency*. New York: Harcourt, 2012.

Kozaryn, Linda D. "Deck of Cards Helps Troops Identify Regime's Most Wanted." *American Forces Press Service*, April 12, 2003. http://archive.defense.gov/news/newsarticle.aspx?id=29113.

Kutz, Christopher. "Causeless Complicity." *Criminal Law and Philosophy* 1, no. 3 (2007): 289–305.

Lin, Patrick. "The Ethics of Autonomous Cars." *The Atlantic*, October 8, 2013. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

Lindley, R. *Autonomy*. Basingstoke: Macmillan, 1986.

Lombard, L. B. "Causes, Enablers and the Counterfactual Analysis." *Philosophical Studies: An International Journal for Philosophy in the Atlantic Tradition* 59, no. 2 (1991): 201–3.

Mackie, Penelope. "Causing, Enabling, and Counterfactual Dependence." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 62, no. 3 (1991): 325–30.

Malle, Bertram. "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots." *Ethics and Information Technology* 18 (2016): 243–56.

Markoff, John. "Arms Guided by Software, Not People, Stirs Fear." *International New York Times*, November 12, 2014.

Mayer, Michael. "The New Killer Drones: Understanding the Strategic Implications of Next-Generation Unmanned Combat Aerial Vehicles." *International Affairs* 91, no. 4 (2015): 765–80.

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

Miller, Jason. "An Ethical Dilemma: When Robot Cars Must Kill, Who Should Pick the Victim." *Robohub*, June 11, 2014. https://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/.

Monroe, Andrew, and Bertram Malle. "Free Will Without Metaphysics." In *Surrounding Free Will*, edited by A. R. Mele, 25–48. New York: Oxford University Press, 2014.

Monroe, Andrew, and Bertram Malle. "From Uncaused Will to Conscious Choice: The Need to Study, not Speculate About People's Folk Concept of Free Will." *Review of Philosophy and Psychology* 1, no. 2 (2010): 211–24.

Moor, James H. "Four Kinds of Ethical Robot." *Philosophy Now* 72, no. 10 (2007): 12–4.

Moor, James H. "Is Ethics Computable?" *Metaphilosophy* 26, no. 12 (1995): 1–12.

Moor, James H. "The Nature, Importance, and Difficulty of Machine Ethics." In *Machine Ethics*, edited by Michael Anderson and Susan Anderson, 13–20. Cambridge: Cambridge University Press, 2011.

Nagel, Thomas. *Equality and Partiality*. New York: Oxford University Press, 1991.

Nussbaum, Martha. *Women and Human Development*. Cambridge: Cambridge University Press, 2000.

Powers, Thomas. "Prospect for a Kantian Machine." *Intelligent Systems, IEEE* 21, no. 4 (2006): 46–51.

Rabin, Robert L. "Enabling Torts." *De Paul Law Review* 49, no. 2 (1999): 435–54.

Scharre, Paul, and Michael Horowitz. *An Introduction to Autonomy in Weapon Systems, Working Paper*. Washington, DC: Center for a New American Security, 2015. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/Ethical-Autonomy-Working-Paper_021015_v02.pdf?mtime=20160906082257&focal=none.

Sharkey, Noel. "Automating Warfare: Lessons Learned from the Drones." *Journal of Law, Information and Science* 21, no. 2 (2011): 140–54.

Sharkey, Noel. "The Evitability of Autonomous Robot Warfare." *International Review of the Red Cross* 94, no. 886 (2012): 787–99.

Sloman, Aaron. "What Sort of Architecture is Required for a Human-Like Agent?" In *Foundations of Rational Agency*, edited by Michael Wooldridge and Anand S. Rao, 35–52. Dordrecht: Kluwer, 1999.

Solan, Lawrence, and John Darley. "Causation, Contribution, and Legal Liability: An Empirical Study." *Law and Contemporary Problems* 64, no. 4 (2001): 265–98.

Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24, no. 1 (2007): 62–77.

Stahl, Bernd. "Can a Computer Adhere to the Categorical Imperative? A Contemplation of the Limits of Transcendental Ethics in IT." In *Cognitive, Emotive, and Ethical Aspects of Decision-Making in Humans and in AI*, edited by I. Smit and G. Laskar, 13–8. Tecumseh, ON: International Institute for Advanced Studies in Systems Research and Cybernetics, 2002.

Teschner, John. "On Drones." *The Iowa Review* 43, no. 1 (2013): 74–81.

Tonkens, Ryan. "A Challenge for Machine Ethics." *Minds and Machines* 19, no. 3 (2009): 421–38.

Veruggio, Gianmarco, Jorge Solis, and Machiel van der Loos. "Roboethics: Ethics Applied to Robotics." *IEEE Robotics Automation Magazine* 18, no. 1 (2011): 21–2.

US Department of Defense. "Unmanned Systems Integrated Roadmap." In *FY2013–2038, Report 14-S*-0553, 1–154. Washington, DC: US Department of Defense, 2013.

Wallach, Wendell. "Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches." *Ethics and Information Technology* 7, no. 3 (2005): 149–55.

Wallach, Wendell, Stan Franklin, and Colin Allen. "A Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents." *Topics in Cognitive Science* 2, no. 3 (2010): 454–85.

Walzer, Michael. *Just and Unjust Wars: A Moral Argument with Historical Arguments*. New York: Basic Books, 2000.

Wright, Richard W. "Causation in Tort Law." *California Law Review* 73, no. 6 (1985): 1735–828.

Zimmerman, Michael. "Moral Responsibility and Ignorance." *Ethics* 107, no. 3 (1997): 410.