Fall 2018

Killing Machines: A Moral Analysis of Autonomous Weapons Systems

Anne Ringelestein

KILLING MACHINES: A MORAL ANALYSIS OF AUTONOMOUS WEAPONS SYSTEMS

by

Anne Ringelestein

A thesis submitted in partial fulfillment of the requirements
for graduation with Honors in the Philosophy

____
Jovana Davidovic
Thesis Mentor

Fall 2018

All requirements for graduation with Honors in the
Philosophy have been completed.

____
Carrie Figdor
Philosophy Honors Advisor

Killing Machines: A Moral Analysis of Autonomous Weapons Systems

By Annie Ringelestein

Abstract: In this paper, I will argue that the use of autonomous weapons systems is immoral for two reasons. The first is that they are unable to form intent when acting – something we need in order to morally evaluate our actions – and that we need to morally evaluate our actions when we carry out actions of moral weight, like taking a life. The second is that autonomous weapons systems cannot account for shifting conventional morality, so they could cause problems in war, where morality can shift very rapidly. I will also contest my view against a number of objections and recount the other literature vital to this debate.

War is a constant through human history. Human beings have always created weapons to fight one another, each more advanced than the next. From the longsword and bow to the M16 assault rifle and nuclear warheads, weaponry has become increasingly lethal. Our modern society now faces an even more advanced weapon: the autonomous weapons system. Similar in appearance to a drone, the autonomous weapons system (hereafter, AWS) is able to select and engage with targets without a human being to make its decisions. The possibility and development of AWS gives rise to ethical dilemmas surround the morality of using artificial intelligence in war, and questions surrounding what it means to be a combatant and if autonomous, yet possibly non-moral, agents can carry out lethal orders. In this debate, there are a number of pragmatic arguments that focus on how AWS might malfunction rather than putting forth principled arguments as to why they are morally flawed, which is what I set out to do in this paper. A number of scholars have already given strong, principled arguments that posit the use of AWS in any capacity is immoral for a number of different reasons. However, there are those on the other side of the aisle that suggest there is nothing principally wrong with AWS, but we ought not to implement them due to a number of pragmatic or normative concerns. Some even hold that, unless something has proven to be pragmatically flawed with AWS, they are permissible to use in war. In this paper, I will suggest that AWS are unethical in principle due to their not being able to form intention before acting, let alone the right one, and we need intention to evaluate the morality of any given action. I will also suggest that AWS cannot account for how quickly facts on the ground can change during armed conflict and would, therefore, be dangerous to implement. In section 1 of this paper, I will define terms key to this debate. I will set the stage by reviewing the other literature on this topic and offer my thoughts on each of the arguments put forth by these scholars in section 2. In section 3, I will argue for my thesis that

AWS are incapable of forming intention, and we are, therefore, unable to evaluate their actions morally, which is required in order to carry out morally weighty acts like taking a life. I will address objections to this argument in section 4. Section 5 will cover my second argument that AWS cannot account for shifting morality as quickly as we would need them to for them to morally permissible to use in war; objections will be addressed in section 6. Finally, I will show in section 7 how my argument is unique from those already put forth by scholars in this field.

I: Terminology

A 2012 directive by the United States Department of Defense defines autonomous weapons system as follows:

> A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes human supervised autonomous weapons systems that are designed to allow human operators to override operation of the weapon system, but can select and engage targets without further human input after activation[1].

In this paper, I will only be focusing on devices that are autonomous in the true sense of the word, meaning that a human operator cannot override operation of the weapon system, even if they wanted to. The definition given by the Department of Defense employs a very narrow sense of autonomy, limiting it only to performing the two tasks at hand (selecting and engaging targets) without the option for human intervention. These sorts of autonomous weapons system are known as weapons that keep human operators “out of the loop”.

The definition of a weapon with a human out of the loop is as follows:

[1] Burri, Susanne. “What Is the Moral Problem with Killer Robots?” in *Who Should Die? The Ethics of Killing in War*, edited by Ryan Jenkins, Michael Robillard, and Bradley J. Strawser, New York City: Oxford University Press, 2018.

> Once activated, the option for human intervention on the actions of autonomous weapons system is impossible. A human being cannot stop the actions of the autonomous weapons system, and the decisions are entirely its own.

This is what I would call a true AWS, because it acts fully independently of its human operators, and humans are incapable of interfering once an AWS had decided what course of action it is going to take.

An AWS can also have a human operator "on the loop" which I would define as follows:

> An autonomous weapons system that has a human operator on the loop is still able to select and engage targets without human operators, but humans are able to intervene and stop the weapon, if they decide the course of action it chooses to take is not ideal.

Weapons with human beings on the loop would be an atypical usage of drones in warfare; these rarely used special drones are able to select and engage targets without human operators, but humans can intervene if they feel they need to. I would suggest that this AWS with a human operator on the loop does not have the same degree of autonomy as AWS with human operators out of the loop, because humans are able to intervene. Critically, this AWS still makes decisions on its own; the only difference is that human beings are able to stop this kind of weapon, should it choose to fire at a child, for example. One of the arguments that I put forth will be sufficient to include both AWS that have humans on the loop and humans out of the loop.

Finally, there are weapons that have human operators entirely "in the loop" which is defined as follows:

> Weapons that have human beings in the loop cannot select and engage targets on their own. They need a human being controlling them to be at all useful. These weapons are, importantly, not autonomous.

A drone that is directly operated by a human could be an example of a weapons system that have human beings in the loop, as it is inactive and unusable without human operators. Typically, when we think of weaponry, this is what we tend to picture.

Philosophical intent has many definitions, but this is the definition I will be working off for the sake of this paper:

> Intention can best be defined as what you plan to do when carrying out an action. It differs from motivation and reasoning because it does not aim to explain why you did what you did; it is merely what you intend to do when acting on a volition.

Importantly, I suggest that intention is key to determining the morality of an action, which I will delve further into in later sections. Ultimately, I will conclude that AWS are incapable of forming intent in the same way we are, so we are unable to gauge the morality of their actions which is necessary to carry out morally weighty actions, so we ought not to use them.

II: Literature

Many scholars have written on this topic, and in this section, I will summarize the arguments that they put forth about the morality of implementing autonomous weapons systems. Robert Sparrow, Duncan Purves, Ryan Jenkins, and Bradley J. Strawser all put forth arguments that suggest the use of AWS is principally unethical, which is the same conclusion that I will be reaching. However, there has not been a lot of work done analyzing AWS from the lens that I am, and my argument is unique compared to all of these other scholars. There are two other scholars – namely, Michael Robillard and Susanne Burri – who suggest that there is nothing principally wrong with these weapons, but there may be other, pragmatic reasons not to use them.

Robert Sparrow in his 2007 article *Killer Robots* puts forth an argument in which he concludes that we ought not to use AWS, because there would be a 'responsibility gap' if the AWS were to malfunction and target a civilian or, even worse, commit a war crime. Sparrow suggests that, if we do not have someone to hold responsible if something like this were to happen, we ought not to implement them in warfare, and to do so would treat our enemy as vermin. His argument, as summarized by Purves, Jenkins, and Strawser, runs as follows[2]:

1. "Waging war requires that we are able to justly hold someone morally responsible for the deaths of enemy combatants that we cause.
2. Neither the programmer of an AWS nor its commanding officer could justly be held morally responsible for the deaths of enemy combatants caused by AWS.
3. We could not, as a matter of conceptual possibility, hold an AWS *itself* morally responsible for its actions, including its actions that cause the deaths of enemy combatants.
4. There are no other plausible candidates for whom we might hold morally responsible for the deaths of enemy combatants caused by AWS.
5. Therefore, there is no one whom we may justly hold responsible for the deaths of enemy combatants caused by AWS.

[2] Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy* 24 (2007). & Purves, Duncan, Jenkins, Ryan, and Strawser, Bradley J. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18 (2015).

6. Therefore, it is impermissible to wage war through the use of AWS. To do so would be to 'treat our enemy like vermin, as though they may be exterminated without moral regard at all."

Sparrow's argument rests on the claim that the ability to be held responsible for one's actions is necessary for moral responsibility, and because one cannot hold AWS accountable for their mistakes, that they cannot be assessed for moral responsibility, so we should not implement them in war. Sparrow holds that the ability to receive punishment is crucial to carrying ought morally weighty actions. Part of my argument sounds similar to Sparrow's, so I will distinguish them later in this paper.

Duncan Purves, Ryan Jenkins, and Bradley J. Strawser (hereafter, Purves *et al.*) also put forth arguments suggesting that there is something principally wrong with AWS. They give two arguments in their 2015 paper *Autonomous Machines, Moral Judgment, and Acting for the Right Reasons*: the anti-codfiability argument and an argument about acting from the right reasons.

Their anti-codifiability argument suggests that moral theory is something that cannot be codified into machines, and an algorithm could never capture the true moral processes that we, as moral agents, use when we make a decision to act. Purves *et al.* cite Hubert Dreyfus who writes that "programmed behavior is either arbitrary or strictly rule-like. Therefore, in confronting a new usage a machine must either treat it as a clear case falling under the rules, or take a blind stab."[3] They suggest that our moral deliberation is neither rule-like nor arbitrary, rather that it is a much more sophisticated process. Furthermore, they hold that programmed behavior could never replicate the moral processes that we use in making decisions.[4] A human being, when confronted

[3] Purves, Duncan, Jenkins, Ryan, and Strawser, Bradley J. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18 (2015).

[4] Purves, Duncan, Jenkins, Ryan, and Strawser, Bradley J. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18 (2015).

with a new scenario, would be able to deliberate about what to do in a way that a machine simply never could, because of the way that they are programmed. However, they suggest another argument, should the anti-codifiability thesis prove to be incorrect, but I am inclined to agree with them here.

The second argument that Purves *et al.* give is that AWS weapons systems are incapable of acting from the right reasons, which would make them morally deficient in comparison to human agents. They hold that AWS mimic moral behavior, but they do not go through the moral processes that a true moral agent would in reaching a decision on how to act. AWS do not choose to act morally; they only follow what their programming says. AI, therefore, cannot act for the right reasons, which Purves *et al.* suggest is necessary to be a moral agent. This argument from the right reasons is similar to one I will put forth later in this paper, so I am inclined to agree with Purves *et al.* and their moral analysis of AWS. I will discuss the differences between my argument and theirs in section seven of this paper, because there may be some initial similarities, but my argument is ultimately unique from theirs.

Unlike the scholars previously discussed, Michael Robillard and Susanne Burri do not believe that there is anything, in principle, wrong with AWS though they both agree that there might be enough pragmatic concern to suggest that they ought not be used.

Robillard suggests that those who suggest there is something principally wrong with AWS have a fundamental incorrect conception of the decision making process an AWS would carry out. He gives 2 suggestions for a better way to think about AWS and their nature: as a "socially-constructed institution that has been physically instantiated"[5] or as a moral agent. If AWS are just socially-constructed institutions, then we would evaluate them morally in the same

[5] Robillard, Michael. "No Such Thing as Killer Robots." *Journal of Applied Philosophy* (2017).

way we would evaluate collective action. If AWS were genuine moral agents, then they would be responsibility-bearers, but they would also, necessarily, have rights and interests. Importantly, Robillard's thesis rests on the claim that there are things other than human beings that can be held morally responsible for their actions. Robillard suggests that we can morally evaluate groups of people for their actions, even though groups of people are not typically thought of as moral agents. If we can evaluate the morality of actions of agents we traditionally take to be non-moral (like large groups of people), then we can evaluate the morality of actions carried out by AWS, which would make them genuine agents rather than non-moral agents. This is an objection I will address later in this paper, as this directly poses a threat to one of my arguments.

Finally, in her article "*What is the Moral Problem with Killer Robots?"* Susanne Burri discusses her qualms with the arguments given by Sparrow and Purves *et al.* Ultimately, she suggests that all of these arguments lack force, and that there is nothing, in principle, wrong with AWS.[6] She argues, in opposition to Purves *et al.*, that robots need not be capable of moral deliberation to resist certain immoral orders; for example, they could be programmed to not fire at children. In response to their argument from right reasons, Burri suggests that the reasons for our actions only truly matter if the action is performed by an agent who is capable of acting on reasons. As AWS are incapable of doing so, they are not subject to this kind of moral evaluation. She compares the critiquing of AWS as morally deficient because they are incapable of acting on right reasons when they are not agents that act on reasons at all akin to "calling a bicycle deficient, because it does not run on four steady wheels."[7] Fundamentally, she claims it does not

[6] Burri, Susanne. "What Is the Moral Problem with Killer Robots?" in *Who Should Die? The Ethics of Killing in War*, edited by Ryan Jenkins, Michael Robillard, and Bradley J. Strawser, New York City: Oxford University Press, 2018.

[7] Burri, Susanne. "What Is the Moral Problem with Killer Robots?" in *Who Should Die? The Ethics of Killing in War*, edited by Ryan Jenkins, Michael Robillard, and Bradley J. Strawser, New York City: Oxford University Press, 2018.

make sense to subject AWS to this kind of moral analysis, because they are not agents that act on reasons. If they were to be agents that act on reasons, then the reasons the agents choose to act on would be important, she concedes. However, because AWS are not agents than can act on reason, it does not make sense to call them morally deficient for not acting on the *right* reasons.

In response to Sparrow, Burri suggests that there are ways of punishing a robot that do not involve making them suffer, so we would not need to worry about the responsibility gap. She also argues for her own position that our duty to protect our own combatants outweighs the moral qualms against AWS, so we ought to develop the technology further. I will address the objection that Burri poses to Purves *et al.* later in this paper as well.

III: Argument from Lack of Intention

Intention plays a key role in determining the morality of our actions, as it is a natural explanation for why we choose to act. We do things because of our intentions; they are not formed to merely explain our actions. To elucidate the importance of our intention in determining the morality of an action, take the example of killing in cold blood versus killing in self-defense. If one were to kill another, only because they enjoy the feeling of physically dominating someone else and taking away another agent's chance at life, then all would agree he is not morally justified in taking a life. However if one were to kill another out of self-defense – someone is attacking someone else, and in order to stop the attack, the victim kills her assailant, for example – it seems intuitive to suggest that this person is morally justified in taking the life. Both of these scenarios have the exact same outcome – one person is dead and the other alive – but they are so morally different strictly because of the intention the killers had when they acted. The person who killed in cold blood had the intention to violently end a life for no reason; the person who killed in self-defense had the intention to save her own life, but this required her to end the life of another, which made the important distinction between these actions. This distinction does not matter just morally; it also matters legally. The person who killed in self-defense likely would not face criminal prosecution, whereas the person who killed in cold blood definitely would. All of this stems from the intention they had when carrying out the act, so intentions must play an important role in determining the objective morality of an act.

Consequentialists hold that morality ought to be determined strictly by the consequences of an action, rather than intention, so those who accept this doctrine are bound to find issues, immediately, with my proposal. In order to show that intentions can matter, even to a consequentialist, I will put forth an example of two shoplifting women. Woman A steals a loaf of bread purely because she enjoys the high that she gets when she takes something that does not belong to her; she is a kleptomaniac. She steals only for the adrenaline rush. Woman B, on the other hand, is stealing the same loaf of bread, because her family is starving, and she cannot afford to make ends meet. If she did not steal the loaf of bread, her family would die. At a minimum, almost anybody would agree that Woman B had more moral justification to steal than Woman A did, because her stealing was for the greater good of her children not dying. Woman A, on the other hand, stole merely to satisfy an illegal desire, so she is not justified in stealing. Consequences, then, can play a huge part in forming intention, because Woman B knew that if she did not steal the bread, then her family would die. Her intention to steal the bread for her starving children, in turn, made her action morally acceptable, because the intention was good and aimed at helping others, rather than just stealing for the rush of it. If consequences play a key role in forming intention and intention is the metric with which we evaluate actions, then a consequentialist might be able to accept some form of dualism which suggests that both consequences *and* intention are necessary to evaluate moral action. If intention did not matter at all, then both the shoplifting example and the example I gave above with the killing in cold blood versus self-defense are null, because the consequences are the same – the bread is stolen and the person is killed. This does not sit well with anybody, as nobody wants to say the person killing in cold blood versus the person killing in self-defense are equally justified in their acts. Similarly, none would suggest that the kleptomaniac is as justified as the needy mother, because of the stark difference in intention. I do not intend to solve the centuries of debate between consequentialism and deontology, rather this is merely to show that talk of intention has its place in consequentialist analyses as well.

Individual human beings are not the only ones who are able to act on intent when they undergo a certain endeavor. In fact, I would suggest groups of people are able to act on intent, so we are able to morally evaluate groups of people based on the intent of their actions. Nations, for example, are able to act on intent when they decide to go to war with one another. Take the following example of war between Nation A and Nation B: Nation A claims that they want to go to war with Nation B to prevent widespread human rights abuses right after Nation A learns about Nation B's vast oil supply. If Nation A really went to war to end the human rights abuses first and foremost, without a thought of the oil in mind, then I would suggest Nation A is justified in going to war with Nation B, and it would be considered a humanitarian war of intervention. However, if Nation A only claims to be going to war to end the human rights abuses, but is really going to steal Nation B's resources, then they are not justified in going to war, and that war would be an unjustified war of aggression. This demonstrates that individual humans are not the only moral agents that can have their actions evaluated based on intention behind them. Nations are, obviously, not persons, but we are still able to determine what the intention was behind going to war, and this intention can change the entire war from being justified to unjustified. I would, therefore, agree with Robillard that certain non-human agents can be considered moral agents, but I believe that he might take it one step too far. This issue will be further addressed in the next section of this paper.

I have demonstrated that intention can make or break the morality of an action, even if the outcome of said action is the exact same in both scenarios. However, intention is also of utmost importance to the laws of war (*jus in bello*). *Jus in bello* is built on the principle of discrimination, which holds that there is a fundamental difference between combatants and non-combatants in war and that combatants ought not engage with non-combatants. Discrimination is ultimately predicated by intention, which would suggest that *jus in bello* rests, at least partially, on the importance of intention. To demonstrate this, consider the example of the tactical bomber vs. the terror bomber.

> **Tactical bomber:** The tactical bomber is given an assignment to bomb a weapons manufacturing plant to destroy an enemy's capacity to produce weapons. The tactical bomber knows that five non-combatants will die as a result of the bombing, but that is not is reason for acting. He does not want to kill the five non-combatants, but their deaths are an inevitable byproduct of his necessary actions.

Compare this example to the terror bomber.

> **Terror bomber:** The terror bomber drops a bomb on the group of five non-combatants, killing them, in order to terrorize the employees of the same aforementioned weapons plant into not going to work. This will secure the same tactical advantage as the first bomber.

I would suggest, and international law agrees, that only the tactical bomber is justified in carrying out the order, despite the fact that they both have the same mission – halting production at a weapons plant. Though they are both given the same order, only the tactical bomber is justified in carrying out the order due to the intention, so intention must be important enough to determine whether or not an act is justified. As stated earlier, *jus in bello* rules rely exactly on this distinction between combatants and non-combatants. Intention is very important in policy implementation, because the only differentiating factor between these two scenarios is precisely their intention. If intention did not matter, the terror bomber would be just as justified in carrying out the order as the tactical bomber, and that does not sit well with most of us. Therefore, intention is incredibly important to determine the morality of an action, and it is impossible to determine objective morality of an act without knowing the intention behind it. The very notion that we are resting an entire field of law on intention of an act demonstrates the importance of it in our actions.

What about beings that, by nature, act with no intention whatsoever? Can we morally evaluate them based solely on the consequences of their actions? Take a philosophical zombie, for example. This zombie is not what we think of when we hear the term; they do not wander around eating human beings, rather they just aimlessly walk, acting for no reasons at all. These zombies have no brain function whatsoever, save for basic instinct in the brain stem that enables them to walk and move. When zombies kill, they necessarily do so for no reason. It is, therefore, impossible for us to determine whether or not a zombie was justified in killing, because they had no intention or reason to do what they did. I would suggest, then, if there is no justifiable reason, that a zombie is not morally permitted to kill, and I feel this is a common intuition. Zombies have no moral justification to kill whatsoever, because of their lack of intention behind their actions, which, in my view, would classify one as a non-moral agent. Because of this, zombies would also not be able to fight in self-defense due to them being unable to form the intention to do so.

Non-moral agents, then, are not morally justified in taking the lives of moral agents under any circumstances, because they are not of the same moral status.

It seems like intentions matter in our moral assessments. I have established that we need intentions to morally evaluate actions, and it follows that better intention equates to – all things being equal – better morality. The better intention we have, the more morally justified we are in carrying out an action, which can be demonstrated in the self-defense example, the shoplifting example, and the terror bomber vs. tactical bomber example. All of these showed the importance of our intention in determining the objective morality of an action, as the cold-blooded killer, the kleptomaniac, and the terror bomber were not justified in acting, while the self-defender, the needy mother, and the tactical bomber were. The only difference between these groups were precisely their intentions in acting, so intentions must be important enough to gauge the morality of our acts. The zombie example also demonstrated that intentions matter for non-moral agents as well, because the zombie – a non-moral agent – is not morally justified in killing specifically because they are not able to form intention to do so. We need to be able to form intention in order to be a moral agent and carry out morally weighty actions, because without intention, it is impossible to determine the morality of an action, and we need to be able to morally evaluate actions in war.

Autonomous weapons systems, in their very essence, cannot act on intention in the same way that we can, because they are just driven by programming, which I could easily compare to the instinct that the motivates the zombie to act. Programming, like instinct, is not the appropriate sort of intention to be able to determine the morality of an action, because programming does not enable AWS to form intention. AWS have a plan and act on that plan, but they are fundamentally unable to form intention in the same way that we are. As Dreyfus claims in Purves *et al.*, “programmed behavior is either strictly rule-like or arbitrary in nature.”[8] The intention that we form as moral agents is rarely this simple, and our intention can incorporate a number of other values that AWS do not have access to – like love or empathy. If our moral processing were as simple as the AWS, then maxims like “it is wrong to kill” or “it is wrong to steal” would be universalized, which would leave our killer in self-defense and our needy mother in the morally unjustified camp, even if this seems counterintuitive. This total reliance on rule-like programming is where AWS critically differs from us and our ability to form intention. As human beings, we are able to take things like love, empathy, and mercy into account when forming intention; these values are alien to AWS and are things they will necessarily be unable to possess. While AWS will always be more accurate in targeting, I hold that values we have as human beings matter in making morally weighty decisions. A mother, for example, might form the intention to kill to protect her child out of love; soldiers at the end of a hard-fought battle might be able to show the opposing side some mercy where it is due. These things are impossible with AWS, and these important values do not play a part in their rule-like plans of action. AWS intention formation, then, is critically different from ours, because they are unable to take outside values into account, so their ‘intent’ will be programmed, rule-like behaviors. Our intent-formation, on the other hand, can take our human values into account as well as the situation on the ground, so we are true moral agents, whereas AWS are not.

Even if an autonomous weapons system was able to explain how they reached a decision to act, there will never be an *intention* for their actions that extends beyond that is what they are programmed to do, which is not the kind of explanation we are looking for. This sort of explanation could explain the plan formed by AWS – i.e. the AWS fired because a combatant-

[8] Purves, Duncan, Jenkins, Ryan, and Strawser, Bradley J. “Autonomous Machines, Moral Judgment, and Acting for the Right Reasons.” *Ethical Theory and Moral Practice* 18 (2015).

age male was holding a suspicious-looking package – but it does not cover intention, which is something completely different. As we discussed above, intent is what someone intends to do when acting, and I will concede that AWS form plans, but they do not form intent the same way we do, as we can take values into account where they cannot. Similarly, if the zombie were able to answer us when we asked why it killed an innocent person, all it would be able to say is that is what it was instinctually driven to do, which does not feel like a very satisfying explanation of intent. While there will be a mere explanation for an action, there will not be an explanation for intent behind the action, so we would be unable to calculate whether or not the action was morally permitted, and we need to be able to morally evaluate actions when carrying ought a morally weighty act like taking a life, so we ought not use these weapons at all.

IV: Objections to Argument I

At this point, it might be objected that if we cannot determine the morality of actions carried out by AWS, then we also cannot determine the morality of actions of large groups of people, because it is impossible to explain intent for a group. Essentially, the question can be asked: what feature do AWS lack and how are they different from a group of senators reaching a decision?

To this, I would suggest that while it might be impossible to determine the intention of an entire group, it is certainly possible to determine intention of individual members of that group and extrapolate the morality of their intention to judge the morality of the entire group. For example, in the war versus Nation A and B scenario, I can say that going to war with Nation B purely for oil extraction rather than humanitarian intervention is immoral, despite that a large group of people decided to go to war rather than just one person.

In a true AWS, the “intent-forming” happens in a closed loop, so human beings – moral agents – are unable to have any input in this formation; we could only sit back and watch. Similarly, human beings are unable to stop the AWS from carrying out an objectively immoral at (killing a child, i.e.) even if they wanted to due to this closed system. A group of legislators, on the other hand, are part of the decision-making loop, so we can track the individual reasons that this certain group of legislators decided to vote in a certain way.

Similarly, when the Supreme Court makes an objectively immoral decision – like suggesting that white people and people of color can be “separate but equal” – we are able to look at the reasoning for each justice and why they decided that would be the best way to vote. This can allow us to evaluate these actions through a moral lens, as there is a clear, understandable intent with these groups of people.

Another objection to this first argument that I have suggested is that we do not need intention to judge the objective morality of an action. For example, killing a child would be objectively immoral no matter what the intention is; we can just intuitively know that there is something horribly wrong with killing a child. While I would agree that this specific example is true, I would suggest that with situations that are not this clear-cut, intentions are required to determine morality.

While killing a child may very well be wrong no matter what the intention is, the tactical bomber vs. the terror bomb really elucidates the importance of intention in our actions. These two were given the exact same order – halt production a weapons plant in order to stop further development on a powerful, game-changing weapon that the other side has – they had *very* different intentions behind carrying out the order. These intentions were so different that it made the tactical bomber morally justified in carrying out the action, whereas the terror bomber is definitely not morally justified, solely because of the intention he had behind his actions.

I do not think this objection holds, because it cherry-picks examples that are very easy to determine the morality of without analyzing the intention behind the action. Most can agree – *a priori* – that torture is a moral evil and ought never be employed; we do not need intention to determine this. However, in more complicated situations, like the terror and tactical bomber or the shoplifting woman, intentions can make a world of a difference and make an act a war crime or just simply carrying out an unpleasant order. Neither of these objections cause fatal flaws to my argument.

Finally, in her paper *What Is the Moral Problem with Killer Robots?,* Susanne Burri gives a number of objections to arguments in the AWS debate. In response to Purves *et al.* right reasons argument, Burri suggests that *right* reasons for acting matter if the agent doing the acting is an agent who is capable of acting on reasons at all. However, she argues, if an action was carried out by an agent who is incapable of acting on reasons, it does not make sense to call that action morally deficient for not acting for the right reasons, if they cannot act on reasons at all. She compares it to "calling a bicycle deficient because it does not run on four steady wheels"[9].

To Burri, I would suggest that acting for the *right* reasons does not necessarily matter, but I believe it is essential to act with intention at all – be it the right intention or not. I have suggested that intention is required to morally evaluate our actions, so while an action performed without an intention might not *necessarily* be morally deficient, in my view, it is impossible to gauge whether or not the action was immoral or not, which is dangerous when carrying out

[9] Burri, Susanne. "What Is the Moral Problem with Killer Robots?" in *Who Should Die? The Ethics of Killing in War*, edited by Ryan Jenkins, Michael Robillard, and Bradley J. Strawser, New York City: Oxford University Press, 2018.

actions that require intention to justify them, like killing. I am not so much concerned with right reasons, as I am with the ability to form intention at all. If an agent is not able to form intention in carrying out an act, then it is impossible to morally evaluate that action, which we need to be able to do, especially in a wartime context. AWS do not form the intention that we need to see in order to morally evaluate their actions, so we ought not implement them in the first place.

I would also suggest to Burri that it *is* appropriate to call AWS morally deficient because they are incapable of acting on intention. She holds that it makes no sense to call an agent incapable of acting on reason morally deficient for not acting on the right reasons, which I can understand. If an agent cannot act on reasons at all, why nitpick that they are bad for not acting on the right ones? However, I hold that AWS are morally deficient for not being able to form intention, let alone the right one, so this objection, too, does not hold.

V: The Argument from Anti-Codifiability of Shifting Conventional Morality

Societal changes can clearly affect conventional morality, and this can clearly be seen in the United States, as well as many other countries, I am sure. Homosexuality, for example, used to be thought of by the majority as a heinous sin that ought to be outlawed, and it was. Today, however, most people are ambivalent towards homosexuality and do not care if people around them are engaging in homosexual relationships. More positively, there is even a lot of gay pride today, which many straight people have accepted and support. The conventional morality towards homosexuality has shifted to be more accepting, and the social stigma surrounding it has definitely subsided (not to suggest that it has completely gone away, homophobia is still alive and well but the majority, in the United States at least is accepting of it nowadays).

This does not just apply to homosexuality. The way that women, people of color, disabled people, and transgender people have all experienced the shifting sociocultural perception of their respective groups. The one that has shifted the fastest, at least in my opinion, is the perception of Muslims and Arabs in the United States. Post 9/11, there was (and still definitely is) a lot of islamophobia, because people were terrified after the attacks. However, today, most people in my generation (and perhaps the general population as well) do not maintain the same fearful distrust of Muslims that the population had immediately following the 9/11 attacks, and this is only seventeen years later. Again, I do not mean to suggest that islamophobia is dead, because it definitely is not, but the conventional morality has shifted to be more accepting these days than it was in 2001.

Our conventional morality, then, can shift so quickly, and it is certainly not something that stays static over a long period of time. Human beings are able to respond to change quickly due to our status as human beings; it was evolutionarily necessary for us to be able to respond to changes in our environment. We are also able to cope with changes in society and with things that are wildly different from what we are used to, due to the highly adaptive nature of our brain. Our brain is designed to adapt to change, even if a lot of us may not like change. Artificial intelligence (hereafter, AI), on the other hand, is not able to respond as quickly as we are to the shifting conventional morality, as they are dependent on programming and algorithms. AI and AWS would need to have someone write code to update it's "morality" rather than just being able to adapt in the same way that we are able.

This ability to recognize the quickly shifting morality is important and fundamental to our status as moral agents, because we can recognize when we might need to alter our beliefs due to new information being given to us. It is also essential for moral agents to acknowledge shifting facts about people, because if we were at war with Al Qaeda, for example, being able to identify one as a terrorist would be helpful. But, if we were to learn that the KKK was to blame for whatever action we went to war with Al Qaeda over, we, as moral agents, are able to quickly categorize this new information and put it to use to target the right people. This feature of our moral status is important in wartime contexts, because facts on the ground can change so quickly. One side may agree to a ceasefire or armistice quickly, and we as human beings would be able to process this and end hostilities. AWS, on the other hand, would need someone to terminate the program (which we have already established is impossible once it is activated) or would need someone to quickly write code to accommodate for this change in facts. This is just not a sustainable way to wage war, because it could lead to re-engaging in hostilities, if the AWS were to continue to target and kill combatants after a declared armistice. This would be, obviously, detrimental to peacekeeping efforts and ultimately would lead to extending the war, which all can agree is a bad thing.

The point I am getting at with this is that AI and AWS are not moral agents in the same ways that we are, because they lack the essential characteristics of moral processes that I have demonstrated are essential to be moral agents. Their decision-making process is fundamentally different from our own, and that matters, because we certainly consider ourselves to be moral agents. Our status as moral agents is what enables us to make those tough choices (life or death, for example), because we have all of the necessary tools to weigh the options back and forth in our heads and arrive at the best possible decision. Understanding that conventional morality shifts quickly is essential to our status as moral agents because of our adaptive nature and ability to process change. AWS, on the other hand, are incapable of doing this, so I would suggest that they are not moral agents.

We ought not implement weapons that are completely autonomous, yet not moral agents, because only moral agents should be able to make those big decisions of life vs. death or war vs.

peace. Only moral agents should have the capacity to kill, because they are able to take in all the relevant conditions, including the quickly shifting morality, and make a decision based off those conditions. The AWS, on the other hand, is only able to make decisions based off of its programming and data collection, so it cannot begin to account for shifting morality. In a truly closed AWS, human beings would not be able to input data into the weapon in order to help it target the right people, if a quick change happened like this. True AWS are limited to using whatever data they are able to collect, and if they cannot recognize peace on the ground quickly, then they are a huge liability in wartime context and ought not be used. Furthermore, this only goes to show that AWS are not moral agents in the same way that we are, and reinforces my belief that we should not use non-moral agents as soldiers in our wars. If AWS are not moral agents, then we ought not to give them the capacity to kill those who are moral agents. Therefore, we ought not implement them at all.

VI: Objections to Argument II

Here it may be objected that this is a pragmatic concern rather than a principled reason to believe that AWS are inherently unethical; this argument holds only if we could not design AWS to be more keen to perceiving changes on the ground. I will concede that it could be seen as a pragmatic worry, but I also believe that this fundamental flaw in their design designates their status as a non-moral agent. This status as a non-moral agent is ultimately at the core of my argument, because I do not think we should be sending autonomous beings who are – at the same time – not moral agents into contexts where they will be making life or death decisions.

Their inability to efficiently recognize and respond to change on the ground like this can preclude one from being classified a moral agent, because of the aforementioned arguments. Conversely, AWS do not possess this trait that we have, *and* they are incapable of acting on intention, so AWS are not moral agents. If, theoretically, programmers were able to work around this worry and be able to input data into the AWS, then I would suggest that the AWS is not truly autonomous in the closed-loop sense of the word. A truly, closed AWS would not be able to receive data inputs from human operators. If the AWS is not a truly, closed-loop AWS, then maybe we do not need to worry about whether or not their morality can shift like ours does, but it also critically limits their ability to act autonomously. My first argument holds whether or not the AWS has a truly closed-loop system.

VII: Comparisons to Purves *et al.* and Sparrow

The first argument I put forth – the argument from lack of intentions – is similar to an argument that Purves *et al.* give in their essay *Autonomous Machines, Moral Judgment, and Acting for the Right Reasons* – namely the acting for the right reasons argument against AWS. Purves *et al.* suggest that the decisions made by AWS would be morally deficient in comparison to ours, because they would not be made for the right reasons. They suggest that AI mimic our moral behavior, but cannot take moral considerations into account as a reason for acting. They use the example of a child suffering; AWS cannot be motivated to act morally, because they follow a carried out set of rules when they choose to engage a target.[10]

This argument is similar to the one I have put forth, but it is critically different in a couple regards. (1) I am not suggesting that the decisions made by AWS would *necessarily* be morally deficient in compared to ours. I am suggesting, rather, that AWS do not act on intention, so it would be impossible to morally evaluate their actions. If the AWS were to, hypothetically, target and kill the leader of a terrorist cell, it would not *necessarily* be a morally deficient act, but we would not be able to gauge the morality of the act, because we would not have access to the

[10] Purves, Duncan, Jenkins, Ryan, and Strawser, Bradley J. "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons." *Ethical Theory and Moral Practice* 18 (2015).

AWS's intentions for acting. We would only be able to evaluate the programmed reasons for why it acted: like age, height, gender, build, etc. suggested that person was a terrorist.

(2) This inability to morally evaluate the actions of AWS is what I believe holds the most weight in this issue. To me, if we cannot evaluate their actions morally at all, and we can only base our conception of morality on the consequences of the actions of and AWS, then we should not use them, because I hold that we need to have the ability to evaluate our actions morally in order to carry acts that would be considered morally weighty. Fundamentally, I do not believe that consequentialism can adequately account for morality. For example, think back to the tactical bomber vs. the terror bomber. Both the actions of both bombers had the exact same consequences: 5 dead enemy non-combatants and a halt in production at a weapons plant. However, we know that their intentions are worlds different: the tactical bomber carries out the order strictly to destroy the other nation's advantage, and the deaths of the non-combatants are just a tragic byproduct compared to the terror bomber who carries out the order, so he can terrorize the non-combatants belonging to an ethnic group he hates. If we were to try to determine the morality of these actions solely on the consequences, then these acts would have the exact same net morality; they were both morally justified, because destroying the weapons facility was the main objective. However, the true immorality lies in their intentions, as the terror bomber is unjustified in acting, because he only acted to be able to murder the 5 non-combatants.

My argument differs from theirs in these regards: I do not *necessarily* think that actions carried out by AWS would be morally deficient compared to ours, but we would not be able to morally evaluate them, and I believe this is what holds the most weight in the debate. If we cannot evaluate their actions, then on what grounds are we able to implement them in war? We are able to question human combatants to get to the bottom of why they acted the way that they did, but we will never be able to do this in the same regard with AWS.

My argument is also slightly similar to that of Robert Sparrow in his *Killer Robots* article, because I am suggesting that we need to be able to morally evaluate our actions in order to carry out actions of moral weight, like taking a life. Sparrow holds that in order to be able to carry out this kind of action, one must be able to be held morally responsible for their actions. While these may initially sound similar, there is one crucial difference between my argument and his.

Sparrow's believes that moral responsibility is required in order to carry out morally weighty acts, because he believes that we must be able to punish someone, should something go wrong. On this view, someone must be held morally responsible for the deaths in war, which is not quite what I am arguing. I do not suggest that AWS need to be held morally responsible, though I agree with him that not being able to punish them is a crucial flaw in their nature. My argument revolves more around intention than responsibility; I have a problem with AWS, because they are unable to form any sort of intent, so we cannot even morally evaluate their actions. If we cannot morally evaluate their actions, then we cannot even get to the point of punishment, so our arguments differ in this respect.

VIII: Conclusion

In summation, autonomous weapons systems should not be employed in a wartime context ultimately due to their status as non-moral agents. I have shown that intentions matter in determining the morality of an action, and that AWS are incapable of forming intentions in the same ways that we are. Human beings are moral agents, due to their ability to form intent and judge their actions based on said intent; AWS are non-moral agents, because they are unable to do so. I have also suggested that AWS cannot account for shifting morality, which is something we are able to do as moral agents – adding another reason to suggest that they are not moral agents and should not be engaged in life or death decisions. This debate is by no means over, and it will only grow in scope once we get closer to the realization of this technology, but I have put forth strong arguments suggesting that AWS are unethical to use due to their lack of intent formation, inability to account for shifting change on the ground, and their status as non-moral agents, due to both of those aforementioned reasons. As this technology grows, it is important to be mindful of the consequences it can have, even if the AWS functions exactly as it was designed to function. Ultimately, I suggest that non-moral autonomous agents are not the type of agents that should be engaging in life or death decisions on the battlefield and to implement them would be an immoral act in and of itself.