The Soldier's Share:

Considering narrow responsibility for lethal autonomous weapons[1]

Kevin Schieman
University of Notre Dame

Abstract: Robert Sparrow (among others) claims that if an autonomous weapon were to commit a war crime, it would cause harm for which no one could reasonably be blamed. Since no one would bear responsibility for the soldier's share of killing in such cases, he argues that they would necessarily violate the requirements of *jus in bello*, and should be prohibited by international law. I argue this view is mistaken and that our moral understanding of war is sufficient to determine blame for any wrongful killing done by autonomous weapons. Analyzing moral responsibility for autonomous weapons starts by recognizing that although they are capable of causing moral consequences, they are neither praiseworthy nor blameworthy in the moral sense. As such, their military role is that of a tool, albeit a rather sophisticated one, and responsibility for their use is roughly analogous to that of existing "smart" weapons. There will likely be some difficulty in managing these systems as they become more intelligent and more prone to unpredicted behavior, but the moral notion of shared responsibility and the legal notion of command responsibility are sufficient to locate responsibility for their use.

- **Keywords:** Artificial moral agents, Justin Watt, just war theory, lethal autonomous weapons, machine ethics, moral responsibility, operational morality, Steven Green, Robert Sparrow

In recent years the international community has seen a number of arguments for preemptively banning lethal autonomous weapons. One argument rests on the claim that such machines, no matter how competent, would cause death for which no one could reasonably be blamed. In particular, Robert Sparrow argues that even if autonomous weapons were capable of attending to complex moral judgments about discrimination, proportionality, and military necessity, no one could reasonably be blamed in the inevitable event that one were to kill unjustly (Sparrow).[2] Put otherwise, in the event that an autonomous weapon were to commit a

[1] I am grateful to the many people that contributed to this essay. In particular, I want to thank Don Howard, Laura Callahan, Heather Roff, Chelsie Greenlee, Richard Schoonhoven, and Kevin Cutright each of whom contributed greatly to the analysis offered in this paper. I also want to thank the two anonymous reviewers for their thoughtful feedback on a previous draft; the essay has benefitted significantly from their insight.

[2] Some have also argued that lethal autonomous weapons, at least as they exist today, should be prohibited because they are limited in their ability to discriminate combatants from non-combatants and struggle to make difficult contextual judgments about proportionality and military necessity (Sharkey). Others have pointed out that there is a

war crime, none among the engineers that designed the machine, the military commanders that ordered it into combat, or the machine itself could justly be held accountable for its misbehavior. As a consequence of this alleged moral responsibility gap, Sparrow claims that no autonomous weapon could satisfy the provisions of *jus in bello* insofar as they require that someone be properly held responsible for unjustified killing (Sparrow). This argument and others like it have become a staple in the literature on autonomous weapons (among others, see Asaro; H. M. Roff; "Mind the Gap"; Gerdes; Taylor)

Granted, there is an obvious sense in which human beings will bear responsibility for the things lethal robots do should we loose them on the world, but Sparrow's concern is a much narrower sense of moral responsibility. Whom, we might ask, is responsible for the soldier's share of killing when machines make the final, fateful decisions about who lives and who dies? Were it the case that no one were morally responsible for killing, even in just this narrow sense, I think the challenge would be significant. However, here I argue that our moral understanding of war, including the legal notion of command responsibility, provides sufficient basis to assess blame for wrongful killing, even where difficult cases arise, and especially in the case of sophisticated machine weapons. Still, I think that Sparrow's intuition has been vindicated in an important sense. While I do not believe that there is a moral responsibility gap, I do believe that an analysis of moral responsibility for autonomous weapons points to something deeply troubling. When it comes to the killing these machines will do in war we will share responsibility, much as we always have, but the burdens of foresight and competence imposed by new and sophisticated technologies will be even heavier than they have been in the past. Far

genuine possibility that autonomous machines might actually prove better at making moral decisions about killing in war than their human counterparts (Arkin). This is a serious empirical question for autonomous weapons policy, but it is not my concern for present purposes. My focus is on questions of moral responsibility, which would remain, even if lethal autonomous weapons proved empirically capable of making reliable judgments about the moral aspects of soldiering.

scarier than a world in which no one could be held accountable is one in which many of us could come to shoulder the blame for no less than the wrongful killing done by machines acting on our behalf.

Addressing moral responsibility for autonomous weapons requires a specification of the types of machines that are of concern for present purposes, and then an account of how those machines fit within our understanding of war as a collective activity; this paper addresses each of these concerns in sequence. The first section presents Sparrow's own assumptions about the foreseeable limits of machine agency. The autonomous systems of greatest concern for Sparrow and which are the focus of this analysis, are those that despite significant abilities to operate free from human control, are limited in their capacity for anything approaching moral understanding.[3] The second section locates such machines within existing systems of accountability that follow from our moral understanding of war and which are reflected in military organization and practice. Even with human soldiers, the collective nature of war is such that the individuals charged with the actual killing do not act towards their own ends, but towards the ends of some political community. Our understanding of war as a collective act provides a basis for understanding how moral responsibility is shared between the soldier, her commander, and in this case, the weapons she uses. The third and final section considers a reasonable objection posed by the liminal case, weapons whose sophisticated behavior makes it particularly difficult to assess their capacity for something approaching genuine moral agency. Even recognizing the difficulties posed by autonomous weapons, it will still be we as humans who bear responsibility for the weapons we use and the harms that they cause.

[3] By "moral understanding," what I have in mind is something like the ability to identify the morally relevant features within a given environment and to weigh those considerations appropriately in making moral choices.

§1. Autonomy, but in what sense?

One of the difficulties of assessing moral responsibility for autonomous weapons is that it is sometimes unclear the types of weapons that we are concerned with or in what sense those weapons might be thought to act autonomously. For example, Sparrow notes that many existing weapons "are capable of acting independently of immediate human control" insofar as they are self-guided, but that such weapons do not raise any new ethical questions, so long as a human operator is selecting the targets (Sparrow 65). Instead, Sparrow's concern lies with the potential next generation of intelligent weapons, those that will be capable of making their own decisions about which targets to engage or how best to engage them (Sparrow 65). This next generation of smart weapons, according to Sparrow, would introduce new and difficult questions for military ethics on account of their cognitive sophistication. He explains that these machines would have internal states roughly commensurable with beliefs, values, and possibly desires and, as a result, would be prone to unpredictable and potentially morally harmful behaviors. While I agree that these machines would present difficult questions, I disagree that those questions are especially new from the perspective of evaluating responsibility for wrongful killing.

It is worth noticing though that much of the confusion surrounding machine autonomy has less to do with the technology than it does a lack of clarity about what it means for a thing to act "autonomously." Heather Roff argues there are two senses in which one might use "autonomy" to describe weapon systems (H. Roff), and I believe the difference between these two senses is helpful in understanding the next generation of intelligent weapons as Sparrow describes them. Roff contends that discussions about autonomous weapons tend to emphasize what she calls an "autonomy-as-law-giving" conception, which traces historically to the Greek notion of *autonomos*, meaning roughly "self-governing." However, she argues "autonomy-as-law-giving is not an appropriate frame for [autonomous weapons] because it is tied to a

particular conception of freedom" which machines, to date, do not possess (H. Roff 14). Whether machines are capable of this sort of autonomy at all is as much a philosophical question as it is an empirical one; the moral and ontological questions surrounding full machine agency run much farther and much deeper than questions about the weapons that we use in fighting our wars—at least for the foreseeable future.

Fortunately, we have at least two good reasons to forestall discussions of weapons that are autonomous in this law-giving sense. First, the technology that will enable full machine agents is not as imminent as the technology that will allow states to populate the battlefield with very capable, albeit morally obtuse weapons. Second, there are a number of philosophical questions attendant to creating fully moral machine agents that are conceptually prior to questions about their role in warfighting. For one thing, if a machine is able to replicate the most human aspects of our experience of the world, it should introduce real questions about our relationship with them (e.g. is a moral machine agent the sort of thing that could possess rights?). While such questions are deeply interesting and are fundamental to our moral understanding of the world and our place in it, I take it that they are not central concerns in the context of autonomous weapons policy. The types of weapons that seem to concern Sparrow may be subject to laws and rules or whatever other stipulations as programmed (or learned), but they are not law-giving in any strong philosophical sense.

Instead, Roff suggests a different conception of autonomy finding inspiration in the Greek notion of *autoexousious*. *Exousia*, she explains, "means *the power* to act, *empowerment or authority*, *authorization or delegation thereof*; rather, it concerns the meaning of a *faculty*." She continues, "for autonomous weapons, debates often circle around whether and to what extent they possess a faculty to carry out the task of selecting and engaging targets" (H. Roff 16). Understood in this way, the questions for autonomous weapons policy are empirical on the one

hand and normative on the other. The empirical question is whether autonomous weapons are capable of carrying out their assigned tasks to some pre-established measure of success (somewhat roughly, can the machines do the things we are asking them to do?). The normative question is whether or not such machines possess the requisite capacities to be so authorized to act (that is, should we allow them to do those things in the first place?). Framed in this way, moral responsibility for autonomous weapons is determined through the process by which we delegate authorities to lethal machines based on the faculties that we take them to have. Moral responsibility though, lies not with the machines to which tasks are delegated, but with those who delegate—those who might be thought autonomous in the self-governing sense and who ought to recognize the moral significance of a machine's limitations. The task for the remainder of this essay is to provide a principled account of how capable, but morally-limited machines fit within an existing system of collective responsibility for violence in war.

§2. Autonomy and command responsibility

In arguing for a moral responsibility gap, Sparrow explains that, "where an agent acts autonomously, then, it is not possible to hold anyone else responsible for its actions. In so far as the agent's actions were its own and stemmed from its own ends, others cannot be held responsible for them" (Sparrow 65). However, if this strict account of autonomy were to hold, then no soldier has ever acted autonomously in war. War is an essentially collective act and although soldiers may act in ways that make them massively blameworthy for unjust harms, their ends are not strictly their own. On the contrary, moral responsibility in war is always shared between those doing the killing and those who direct them to violence. While there is a sense in which this is also true of the political leaders who direct armies to war, there is a much narrower sense in which military commanders bear direct responsibility for the actions of their soldiers.

That is not to say that soldiers do not bear significant responsibility for the choices they make about matters of life and death, but that military commanders are responsible for the manner in which violence is conducted under their command. This is particularly true where there is reason to doubt that those delivering the violence have the means and wherewithal to do so discriminately, be they human or machine. This notion, that of command responsibility, is not entirely unique to the military (at least insofar as parallel concepts apply in other parts of our moral lives), but it is a significant enough aspect of our moral understanding of war to answer Sparrow's challenge. Sparrow's concern to avoid placing unfair blame on military commanders for the choices of autonomous weapons is admirable, but ultimately, I believe, it is mistaken.

This section describes the moral role of the military commander by considering vignettes from recent U.S. military history. By referencing actual historical cases, I hope to demonstrate that my arguments are grounded in something like a commonplace understanding of military ethics. That is important insofar as I am claiming that accommodating autonomous weapons does not require significant revision to moral theory or military practice. The section advances three primary claims. First, that even where human soldiers act very badly, the commander, fairly or not, may bear significant blame for her soldier's misbehavior—the gravity of war is such that the commander's competence, not just her knowledge or intentions, rise to the level of moral concern. Second, it is often the case that a military commander bears direct moral responsibility for the actions of the soldiers within her command because her decisions are pre-requisite on the possibility of necessary and proportionate violence. Lastly, I argue that existing "smart weapons" offer a reasonable precedent for adjudicating moral responsibility for autonomous weapons.

2.1. Anatomy of a war crime

Moral responsibility in war is always shared between the soldiers who do the killing and the commanders who direct them to those ends. Granted, moral responsibility is shared exclusively among moral agents, a point so obvious that it would not bear mentioning in other contexts. However, some autonomous weapons are already capable of performing tasks traditionally reserved for human soldiers in the identification, selection, and engagement of targets. With that in mind, the commander's responsibility for ensuring that violence can be done within the confines of *jus in bello* takes on an even greater importance with robotic weapons. This account starts by analyzing the role of the commander in traditional contexts as a basis for understanding the impact of evolving military technologies.

There is no doubt that individual soldiers often bear moral blame for wrongful killing. To paraphrase Jeff McMahan, being a soldier is not just physically risky, it is also morally risky (McMahan). However, war differs from most aspects of private life in that soldiers may not bear that blame alone. In many cases, even when a soldier is personally responsible for great evil, our intuitive sense is that others may share in the blame. This is clearly evident in the case of Steven Green, whose atrocities in Iraq in 2005 were among the most disturbing in recent memory. Green's blame is obvious, but his chain of command was also morally culpable for failing to limit his duties in ways that reflected overwhelming evidence that he was incapable of killing discriminately—a fact that is surprisingly helpful in understanding moral responsibility for autonomous weapons.

By the time Private First Class Green snuck away from his traffic control point in rural Iraq with three platoon mates, it was clear that he should never have been a soldier in the first place. That was true long before the four soldiers committed the premeditated rape and murder of a fourteen-year-old Iraqi girl named Abeer Qassim Hamzah al-Janabi. In addition to the torturous death the group inflicted on the young girl, Green executed her parents and her six-

year-old sister, trading between his shotgun and the AK-47 he had stolen to stage the killings as sectarian violence. The grisly details notwithstanding, the crime stands apart in its depravity and malice aforethought. Green and his conspirators did not stumble into tragedy, they sought it ought; fueled by contraband Iraqi whiskey, they donned disguises and snuck away from their posts, fully intent on doing great evil.

To suggest that Green should never have been a soldier is in no way to absolve him of blame for what he did. Quoting Jim Frederick, whose book *Blackhearts* details the ugly incident in the violent contexts of the unit's 2005 deployment to Iraq, "nothing can absolve Steven Green (and his co-conspirators) from the personal responsibility that is theirs and theirs alone, for the rape of Abeer Qassim Hamzah al-Janabi, her vicious murder, and the wanton destruction of her family."[4] Whether or not Green was the mastermind of the incident is beyond the point; his racism and sociopathy were the catalyst that turned drunken, violent ideation into a quadruple homicide. But, Frederick continues:

> Leading up to that day, a litany of miscommunications, organizational snafus, lapses in leadership, and ignored warning signs up and down the chain of command all contributed to the creation of an environment where it was possible for such a crime to take place (Frederick 17).

Green was underqualified for enlistment from the start, requiring a "moral waiver" in order to even enter the service and his failing mental health was hardly a secret by the time he shot Abeer al-Janabi and her family. Whatever Green's role, there seems plenty of blame left to go around. Others, particularly his chain of command, were responsible for the fact that Green continued to soldier while his mental health deteriorated. Even if those failings were matters of negligence

[4] The book's name is a double entendre. Along with the gross depravity of the act, Green's unit, the 101st Airborne Division's 2nd Brigade, wears helmets adorned with a small, black heart.

and omission, rather than malice and commission, there is good reason to think they were still failings of a moral sort.

Like most tragedies, the Mahmudiyah Massacre did not occur in isolation and the horrible consequences owe as much to a long sequence of poor decisions as they do the final, fateful ones. Green had a difficult childhood and by the time he tried to enlist, he had been clinically diagnosed with depression and had racked up several criminal convictions for drugs and underage drinking; at one point he spent several weeks in juvenile detention and later, several days in jail (Frederick 90). Tragic as was his upbringing, the criminal record alone was enough to disqualify Green from enlistment. However, in 2005, desperate to fill its ranks with volunteers to fight simultaneous wars in Iraq and Afghanistan, the Army was granting "moral waivers" at a rate almost 50% higher than pre-war levels. The "moral waiver" essentially forgave Green's criminal past for the purposes of his enlistment. Given the nature of the criminal transgressions though, the more troubling oversight lay in failing to address Green's openly racist worldview, which was apparent well before combat stress began contributing to his unraveling.

Whatever the circumstances of his enlistment, it was leaders in Green's unit who were most complicit in the tragedy. The soldiers of the Black Hearts were immersed in violence and death, a fact that goes a long way toward explaining how conditions deteriorated to the point where a soldier like Green was tolerated at all.[5] Still, it is startling to learn that before killing the al-Janabis, Green expressed his homicidal desires to not only his platoonmates, but also to a psychiatrist, the brigade chaplain, and his battalion commander. For a soldier with Green's reputation for racist invectives, it is possible that his peers just never took him seriously, but a

[5] John Diem, a team leader and sergeant in Green's company, described their hellish deployment to Iraq by explaining, "you line up three people in a row, and one of them dies" (Frederick 506).

screening with the combat stress team should have been different. During that screening, Green noted wanting to kill Iraqis four separate times and even listed his interests on the intake form as, "None other than killing Iraqis" (Frederick 175). For whatever reason, Green's direct leadership failed to recognize the grave threat he posed to others. Eventually, Green was diagnosed with a preexisiting antisocial personality disorder and discharged from the service, but by then it was too late. The al-Janabi family had been dead for months. The Army had already failed them and it had already failed Green, whose behavior constituted overwhelming evidence that he was psychologically incapable of killing discriminately—that he was literally incapable of soldiering in any moral sense.[6]

One thing that seems apparent in consideration of the murders is that the blame we attribute to Green stands more or less independent of the blame we attribute to the leadership in his unit. Even if Green had not killed the al-Janabis, we might think his chain of command negligent in its duties. In fact, the one laudable postscript to this shameful incident seems to demonstrate exactly this point. Justin Watt, then a twenty-two-year-old private, learned of the killings indirectly and reported them at great personal risk. It is not an exaggeration to say that Watt, often serving in remote locations and under arms with the murderers, literally risked his life in the hopes of bringing them to justice. Watt explains:

> If I kill someone in combat that's the risk that the other guy involved has agreed to take… But civilians are different. The guys who did this had to pay. Not to say that if I never turned them in, they wouldn't be paying for it in their own heads. Your own conscience is worse than any punishment that anyone else can lay on you… But that's not good enough. Not for that shit. (Frederick 350).

Watt's story is worth remembering on its own merits but suffice it to say that he did all that morality asked of him and probably more. Still, there is ample reason to think at least some

[6] Since the perpetrators were not uncovered for months after the murders, Green was recalled to military service in order to be tried for his central role in the crimes. He was ultimately convicted and sentenced to life in prison without possibility of parole. He died in prison in 2014 two days after he had attempted to hang himself in his cell.

members of Watt's chain of command failed him as well; these are many of the same leaders, mind you, that failed to recognize to the dangers posed by an openly-racist sociopath in their ranks. According to one account, Watt's battalion commander was so upset with gaps in the initial report that even after the story had been corroborated, he remarked "He's lucky I don't take him up on charges for making false official statements!" (Frederick 369).[7] The reaction is particularly discomforting considering that this is the same battalion commander to whom Green had expressed a desire to "kill all Iraqis" before murdering the al-Janabis. The point is that the culpability of the commander stands independent of our attributions of praise and blame in the individual cases; Green's behavior was condemnable beyond words and Watt deserves every bit of our praise, but the chain of command deserves blame in both cases.

2.2. From blame to responsibility

The intuition that Green's chain of command was in some way to blame for the killings is helpful in appreciating how moral responsibility is shared in war. There are two senses in which one might speak of moral responsibility. On the one hand, one might ask "whether a person bears the right relation to her own actions, and their consequences, so as to be properly held accountable for them" (Talbert). This line of inquiry might lead us to question, for example, whether Green's clinically-diagnosed psychopathy should mitigate his blameworthiness. Alternatively, one might speak of moral responsibility in the sense that we say judges, doctors, or military officers have role-specific duties to which they *ought* attend (Eschleman). When we

[7] Considering that Watt pieced together what happened from several second-hand accounts, it is actually impressive that his report was as close to truth as it was. For example, Watt pieced together the fact that there were multiple soldiers involved, despite the fact that the first story he had heard about the incident only implicated Green. Nevertheless, the fact that his battalion commander harbored any animosity towards Watt for coming forward with the information he had, no matter how incomplete, probably says everything one needs to know about the climate in the unit at the time of the murders.

speak of command responsibility, it is this latter sense of responsibility we have in mind and as such, the role-specific duties of commanders follow from our broader understanding of war.[8]

In Green's case, clinically diagnosed antisocial personality disorder (sociopathy or psychopathy, colloquially) left him unable or unwilling to draw any meaningful distinction between combatants and non-combatants. There is a real sense in which he was just blind to obvious features of the moral world. I am not interested in advancing an argument exculpating Green in any sense, but I will point out that his clinical diagnosis suggests a lack of moral sensitivity not entirely dissimilar to the concerns posed by otherwise capable, but morally obtuse autonomous weapons. It is for this reason that his commander bears responsibility for Green's ever having the opportunity to deliver so much evil onto the al-Janabis. Green's commander might not have pulled the trigger, but he should have known that his soldier was grossly incapable of killing in any moral sense. Green was, we might say, a war crime waiting on an opportunity, which time and negligence delivered in sufficient measure.

One might mistakenly assume that command responsibility reflects a sort of indirect responsibility for killing in war. That is, Green was directly responsible for killing the Iraqi family, but his commander's responsibility was indirect in the same sense that I might bear indirect responsibility for a friend's offensive language should I choose not to confront him. I might not be directly responsible for his saying offensive things, but I am responsible for setting the conditions in which he is able to continue to offend. This is a fundamental misreading of command responsibility, both in terms of legal precedent and in terms of our moral understanding of the role of the military commander. Soldiers do not kill on their own behalf, but on behalf of a political community that directs them to violence and bears some

[8] In particular, I have in mind that war is both rule-governed and a collective exercise of political communities.

responsibility for the manner in which it is delivered. Command responsibility is the mechanism by which the political community shapes that violence. It is by virtue of command responsibility that other soldiers in the chain of command are directly responsible for ensuring violence can be carried out discriminately and proportionately.

The American policy of fire-bombing Japanese cities during the Second World War warrants consideration in this regard. The bombings, conducted under the command of General Curtis LeMay, sought to compel an unconditional surrender by burning entire Japanese cities to the ground. Robert McNamara, who later rose to fame as the U.S. Secretary of Defense during the Vietnam War, was the lead campaign planner for LeMay. Reflecting on the bombing campaign for a 2003 Errol Morris documentary film, McNamara observed (and not without a touch of irony), that "killing fifty to ninety percent of the people of sixty-seven Japanese cities and then bombing them with two nuclear bombs is not proportional in the eyes of some people, to the objectives we were trying to achieve" (Morris). But he continues, "LeMay said if we'd lost the war, we'd all have been prosecuted as war criminals. And I think he's right. He, and I'd say I, were behaving as war criminals" (Morris). McNamara's reflections are fascinating for a number of reasons, but maybe none more so than the fact that he acknowledges that the commander was morally and criminally responsible for the ends towards which he set the soldiers in his command. Independent of the individual decisions of each bombardier in his 21st Bomber Command, LeMay was ultimately accountable for considerations of proportionality (to say nothing of the fact that the fire-bombing was wildly indiscriminate). Without personally killing anyone, there is a sense in which LeMay was morally responsible for literally millions of non-combatant deaths. It is also worth noting here that we seem to think LeMay was far more culpable than the bombardiers of the 21st Bomber Command (at least I have never heard anyone

suggest it would have been appropriate to prosecute the bombardiers for the fire bombings). It is worth wondering out loud what difference autonomous, robotic bombers would have made.

What is critically important here is that many of the decisions commanders make are themselves moral decisions and lie beyond the scope of the soldiers that do the killing. McNamara captures this burden of command, recalling LeMay's response to a pilot, despondent over losing his wingman in a low-level bombing run. LeMay, who was rarely emotional, said, "You lost your wingman! It hurts me as much as it does you. I sent him there! And I've been there, I know what it is. But, you lost one wingman and we destroyed Tokyo" (Morris). It is not just that individual soldiers do not decide the objectives in war, but that in many cases, they are in no position to evaluate the proportionate costs and benefits of the killing they do in terms of the broader military and political ends of that violence. In some sense, that is what distinguishes responsibility for killing at the commander's level and at the soldier's level. We might say that the soldier's responsibilities are narrower concerns of discrimination and proportionality, while the commander is concerned with those things, but in a broader sense.

Somewhat beyond the question of moral blame, I think McNamara is correct that LeMay, among others, would have been prosecuted for war crimes had the U.S. lost the war. The unfortunate legal reality that prosecutions for war crimes are so dependent upon the outcomes of wars notwithstanding, the moral responsibility of military commanders is codified in legal precedent. It is worth considering the precedent of international law as it pertains to the moral notion of command responsibility for two reasons. First, without supposing that the law necessarily aims to capture morality as part of its content, we should at least observe the in-principle consistency between international law and the moral notion of command responsibility I have advanced to this point. The consistency of the two notions, I believe, provides indirect support for my argument. Second, the International Law of Armed Conflict directly shapes

military practice for the United States and many other signatories to the Geneva Conventions. As I have argued that autonomous weapons do not demand wholesale revision to military practice, international law is also significant in that regard.

A legal precedent called the Yamashita standard (sometimes also called superior responsibility) holds that commanders are morally and legally responsible for some aspects of the harm their soldiers do. The standard is named for the Japanese general subsequently executed for failing to prevent his soldiers from committing widespread atrocities in Manila and the Philippines during World War II (no doubt reinforcing McNamara's point). Maybe most significant about the Yamashita standard is that it holds a commander responsible for what he *should have known,* rather than what he *did in fact know* (McCaffrey 12). This normative standard of command responsibility is not limited to high command; commanders exercise responsibility for how violence is done at all levels in the military hierarchy. I do not think it an exaggeration to say this is that defining feature of command legally, as much as it is morally.

Beyond the legal precedent established by the Yamashita standard, international law is explicit in establishing the extent to which a commander is responsible for the actions of her subordinates. No less than the Geneva Conventions formalize command responsibility in Additional Protocol I of 1977:

> The fact that a breach of the Conventions… was committed by a subordinate does not absolve his superiors from penal or disciplinary responsibility… if they knew, or had information which should have enabled them to conclude in the circumstances… that he was committing or was going to commit such a breach and if they did not take all feasible measures within their power to prevent or repress the breach (*Protocol I, Geneva Conventions*)

Admittedly, *Protocol I* leaves considerable room in interpreting the legal scope of command responsibility and jurisprudence varies from country to country. For example, the U.S. does not have specific provisions for prosecuting commanders under the Uniform Code of Military

Justice, the Federal law defining criminal conduct for military servicemembers.[9] However, the Department of Defense (DoD) Law of War Manual does establish a more stringent standard for evaluating a commander's criminal liability for war crimes. The manual points out that the law of war presupposes that the exercise of command is the means through which war crimes are to be avoided and even that "one of the requirements for armed forces to receive the privileges of combatant status is that they operate under a responsible command" (*Law of War Manual*). Further, it specifies that the commander's duties extend beyond reporting war crimes and include, among other things, training and education. As a matter of practice, it may be difficult to prosecute war crimes under the provisions of command responsibility, but that is not because legal doctrine fails to capture the scope and significance of the concept in understanding warfare as an activity governed by moral principles.

Both legally and morally, command responsibility consists in ensuring that soldiers are capable of killing within the strictures of the Law of Armed Conflict and then putting them in positions to do so. As a matter of practice, command responsibility is often manifest in the most routine aspects of military life: training, education, attending to health and welfare, documenting performance, counseling soldiers, and maintaining equipment. Despite being routine, each of these activities is important to killing morally in war. Without sufficient training soldiers have little hope of killing effectively, let alone discriminately. Without reliable equipment or healthy soldiers, the same holds true. So it is in the collective exercise of these role-specific duties that a military commander makes judgments about the competence and suitedness of her soldiers to killing morally. It was other soldiers in Green's chain of command who spent the most time with him and should have been most attuned to his shortcomings as a soldier (and as a human being).

[9] This is an oversight that some within the Department of Defense have called for correcting (Walsh).

In this sense, it is not a stretch to say that command responsibility culminates in an individual commander's judgments about the competence and trustworthiness of the soldiers she will ask to kill in war. To the extent that these judgments fail, the commander, even by matter of negligence or incompetence rather than malfeasance, may be to blame for no less than unjust killing. Sparrow suggests that in the context of autonomous weapons this is unfair. I suppose I am inclined to say so much the worse for fairness. Command has always imposed a heavy burden on commanders because they are uniquely able to ensure that the violence being done in our name is being done well. Maybe more importantly, it is in virtue the fact that they are uniquely able to set the conditions for killing to be done within the strictures of *jus in bello* that military commanders assume directly responsible for failing to do so. This is a significant moral burden, of course, but as forever, "Uneasy lies the head that wears a crown" (Shakespeare and Weis).

2.3. Smart weapons and war

Existing "smart" weapons, which already incorporate sophisticated processes to identify and target combat vehicles, provide a useful precedent in the exercise of command responsibility for autonomous weapons. The U.S. Navy's Mark 60 Encapsulated Torpedo ("Captor" for short), which came into service in the late 1970s and has since been retired, was among the first lethal "autonomous" weapons (regardless of how it has been classified in practice). The system was an air-dropped, undersea mine, which when triggered, would launch an acoustically-guided torpedo. The Navy claimed Captor could discriminate the acoustic signature of Russian attack submarines from other maritime traffic (Finney). Rather than speculate about Captor's accuracy in discriminating between Soviet submarines and commercial traffic, it is probably enough to point out that every system has an error rate. This is true of state-of-the-art systems today, it is true of human soldiers, who often misidentify targets, and it was certainly true of Captor.

The sailors who deployed Captor, and maybe more aptly, their military commanders, were morally responsible for the system's use. For starters, whatever the error rate, the Navy was responsible for deploying Captor in locations that minimized the risk to civilian traffic, commensurate with the importance of the military objective. It is not unfair to assume that the acoustic signature of a submarine and an oil tanker are rather different, but it is naïve to assume that the system posed no risk of accidental killing. More to the point, absent other antecedent conditions, "being a Soviet submarine" is not sufficient to warrant an engagement. Obviously, there would need to be some state of political hostility, to say nothing of military necessity in a given case. As revolutionary as the technology was for its time, human beings still exercised moral responsibility over Captor by managing the timing, location, and manner of its employment.

A system like Captor suggests a positive account of moral responsibility for the use of ethically limited systems: where a system is ethically limited, command responsibility for the use of that system includes restricting the contexts of its employment in ways that can be reasonably expected to produce moral outcomes. Doing so requires commanders to recognize what is new and different about autonomous weapons, of course, but equally as important is recognizing how they fit within existing practices. One obvious example is fire control measures, which are every bit as important to mission planning in a world without autonomous weapons as they would be in a world with them. Fire control measures restrict engagements by time, command, or phase of an operation, as well as geographical location and possibly even target type. Critically, these measures are tailored as appropriate to a particular mission, personnel, weapons system, and the commander's risk tolerance both in terms of mission success and collateral harms. These measures are no doubt necessary for safely incorporating autonomous weapons into military operations and the difficulty of doing so does suggest an interesting blurring of the division

between the soldier and her equipment. Traditionally, it has been human soldiers to whom fire control measures were tailored and upon whose judgment and compliance their effectiveness depended. As autonomous weapons become more sophisticated and we begin to trust their judgment more than just their reliability, the more important it will be to acknowledge the ways in which these systems are challenging existing paradigms of command and control. It is not necessarily that we lack the means of controlling these systems, so much as those means have traditionally been applied to human soldiers, rather than their weapons.

Maybe the best way to emphasize this shift is to consider the risks posed by a loss of communications with an autonomous weapons system. Imperfect communications are a condition of modern warfare inasmuch as a problem to be solved through planning. In the event of broken communications with human soldiers, the expectation is that they are able to infer the reasonable courses of action from mission context and a deep understanding of the commander's intent (including the desired end state of an operation and the commander's risk tolerance). This type of initiative requires a level of trust which it would be altogether inappropriate to apply to a system like Captor. It is worth noting here that the ability of human beings to act within the broad latitude of something like a commander's intent highlights the importance of the distinction between human soldiers that may be capable of something approaching autonomy in the "law-giving" sense and the machines to whom we might delegate some more limited permission to act on our behalf.[10] The encapsulated torpedo was only capable of one very narrow interaction within its environment; not to put too fine a point on it, but the autonomous torpedo mine could identify a thing that makes a certain noise and it could kill that thing. As a result, the naval commanders who employed Captor did not have to concern themselves with its

[10] I am appreciative to Kevin Cutright for pointing out that this example emphasizes the importance of the distinction offered by Roff in understanding the moral role of a system like Captor.

values or anything like them; it was enough to trust that the system would operate reliably as advertised (Roff and Danks). This difference does not suggest that military practice lacks the means of accommodating more sophisticated weapons, so much as it suggests that commanders will face a delicate balance between restricting those systems as they might restrict conventional weapons and restricting them as they might restrict human soldiers. Managing that balance, despite the technology, is ultimately just another exercise of command responsibility. Just as with human soldiers it is through the commander's judgment about the competence and trustworthiness of an autonomous weapon—no matter how sophisticated—that she will exercise responsibility for the killing that machine might do.

§3. The liminal case

One might object that this analysis ignores the practical difficulty of adjudicating responsibility for sophisticated robotic weapons. Somewhere between existing weapons and morally-fluent machine weapons, one might argue, are cases for which commanders cannot account and which pose grave risk of unjust killing. There is no doubt that the practical difficulties facing commanders stand to grow much more difficult as machines are capable of greater autonomy and become more ethically sensitive. The more machines become capable of extremely sophisticated, flexible behavior in complex environments, the more difficult it will be to identify their specific moral limitations. Of particular concern are machines that can navigate the world in sophisticated and at least ostensibly intelligent ways, but lack important moral sensitivities, in spite of outward appearances to the contrary. In the context of lethal "autonomous" weapons, these highly-capable, morally-obtuse systems represent the most difficult case for policymakers and the potential worst-case for military commanders. However, command responsibility is an active, rather than a passive feature of our moral understanding of

war. Whatever weapons we use, commanders exercise great power over the specific roles that we allow them to play. In this vein, I want to argue that these cases for lethal autonomous weapons are roughly analogous to those of human teenagers and that the legal precedent of affording teenagers a sort of provisional moral status is likewise instructive for the active exercise of moral responsibility for sophisticated autonomous weapons.

Adolescence is an interesting developmental period in human beings. Typically, in adolescence, children have developed sufficient autonomy that they are less reliant on their parents than at earlier developmental periods. They are able to clean themselves (at least in theory), feed themselves, and in many cases, even drive and work. They are able to take on more substantial roles in society and are correspondingly accountable to their choices. However, there is a very real sense in which adolescence is still a developmental period. Findings in neuroscience show that adolescents are still cognitively limited compared to mature adults; they are comparatively less able to recognize emotional responses in others, to modulate their own emotional responses, are more prone to engage in risky behaviors, they are more susceptible to peer influence, and are at higher risk of negative affect or depression (Pfeifer and Blakemore). None of this should come as any surprise to anyone familiar with teenagers—particularly parents—but, it is important to notice that these behavioral differences are correlated with differences in neurophysiology between adolescents and adults.[11] It is not just that teenagers do not always act as responsible moral agents, despite outward appearances that they might be capable of doing so. They are actually biologically less sensitive to important moral features of

[11] In particular, there appear to be significant differences in patterns of activation in the amygdala (the effective center of emotional processing) and the pre-frontal cortex (which is important in impulse control and conflict resolution, among myriad other things) (Pfeifer and Blakemore).

the world than adults. Legally, this has resulted in laws that tend to grant teenagers a sort of provisional or attenuated status as moral agents.

The evidence that we regard teenagers as limited in their ability to exercise moral agency is substantial. Teenagers are, among others things, differentiated from adults and other children in their legal abilities to work, consent to medical procedures, participate as members of a community (to include operating motor vehicles and voting), their legal liability, and the degree to which they are subject to punishment (Scientific Analysis Corporation et al.). Far more instructive for present purposes, though, are the common principles that underlie this type of limited legal status. At the broadest level, these limitations seem directed towards two ends: protection and development. In the first case, limiting teenagers' legal standing within the community protects both them and others from the worst possible consequences of their underdeveloped capacity as agents. Somewhat tongue-in-cheek, we might say that these laws differentiate teenagers as a way of ensuring we all survive their adolescence. However, this attenuated status also allows teenagers greater permissions than younger children, giving them some access to the decisions and responsibilities that will enable their development into mature human beings—at least that is our hope.

Correspondingly, we can point to two over-arching principles about how we should manage moral responsibility for lethal autonomous systems. First, we should exercise conservatism in estimating the moral abilities of machine agents. Since we ultimately decide how these machines will act on our behalf, we exercise great power in the ways in which we choose to empower them. In the case of "autonomous" weapons, that means circumscribing conditions for their use that are probably much narrower than the machines' capacity to act autonomously (in the sense of being autoexousious). There is already a precedent for this approach in practice.

The U.S. Navy's Aegis Combat System (ACS) is designed to protect naval groups from airborne threats ranging from attack aircraft to ballistic and cruise missiles. The ACS is essentially a suite of technologies that includes long-range radar to monitor airborne traffic, software to prioritize threats based on radar signature and heading (among other factors), and various means of defensive engagement—at longer ranges, Aegis employs various ship-based surface-to-air missiles, and for close-in defense, it employs the Phalanx CIWS (pronounced "sea-wiz" and short for "close-in weapon system"), an imposing, large-caliber Gatling gun. Together, these technologies provide naval groups a comprehensive, layered defense against airborne targets (DeLuca et al.).[12] While the system is capable of fully autonomous defensive engagements as allowed by US policy, that mode is generally reserved for close-in, "last-ditch" defense. For more distant threats, the system is generally operated in control modes whereby the ACS provides targeting recommendations to a human controller, who makes engagement decisions (Deputy Secretary of Defense; Singer). Although the system is capable of fully autonomous naval defense, policy restricts that mode of operation to the most desperate attempts to protect the ship against inbound threats. In those cases, the speed of modern warfare prevents timely human intervention (Hammes) and the risk of misidentifying an inbound threat is less severe; for example, an inbound missile has a very different flight profile in its terminal phase than passing civilian traffic. Despite its conservatism in delegating tasks to the ACS, the policy still entrusts the system with morally significant decisions. Hanging in the balance in close-in engagements are not only the fate of the potential threat, but those of the crew as well. Instead of some more general permission to select and engage targets, the policy places the burden of proof

[12] The ACS also integrates protection against sub-surface threats, but the relative speed of airborne threats makes them much more germane to present purposes.

on those delegating a particular moral decision to a particular machine under a particular range of circumstances.

Second, it is incumbent upon human agents to safeguard against the most potentially catastrophic outcomes. Lethal "autonomous" weapons will make mistakes; they are not immune to that aspect of technological progress. However, it is in the choices that human agents make about when and how to empower smarter machines that we exercise control in preventing the weightiest consequences. At a minimum, lethal "autonomous" weapons employment must consider fail-safes, limit munitions and contexts of employment, and identify standards for safety testing in research and development. It should be needless to say that morally-limited autonomous weapons should not carry nuclear payloads, but less obvious limitations are likely the backbone of responsible policy. There are likely other prudential principles that could be important to maintaining positive control over sophisticated autonomous weapons, but whatever they are, they form the basis of a moral obligation in the specific manner by which commanders choose to exercise responsibility for such weapons. There is no doubt that machines capable of great autonomy with limited moral sensitivity are a distressing possibility, but there should be some comfort in the realization that where there is doubt about the machines, there is unquestionable human moral responsibility for the things they do. Much, though not all, of that responsibility falls on military commanders; fortunately, this is one aspect of war that technology cannot change.

§4. Conclusions

The difficulties posed by lethal autonomous weapons in terms of moral responsibility are not so unique that they fall outside existing systems of accountability or notions of command responsibility. Barring the development of full artificial moral agents, a monumental advance from the current state of the art, humans alone will bear responsibility for the machines and weapons they use in fighting wars—even in the narrow sense that concerns Sparrow. There may ultimately be good reason to refrain from building and deploying autonomous weapons, but it will not be because new technologies lie beyond moral responsibility, at least not in the context of the military and war. There will no doubt be difficult cases, but human beings, individually or in small groups, will bear direct responsibility for the choices of the sophisticated weapons that it seems ever more likely we will eventually field. The upshot of this reality though is that those who bear the most direct responsibility for these advanced weapons, military operators and commanders in the field, have great incentive to exercise conservatism in estimating the abilities of these machines and to go to great lengths to safeguard against the worst consequences of using these weapons imprudently.

In concluding, I would like to emphasize two points. First, Sparrow makes the mistake of considering autonomous weapons technology in isolation from its techno-social contexts. Technologies play an important role in human life (morally and otherwise), but it is often a mistake to try to understand them in isolation from the specific manner in which we choose to use them (Vallor). Second, concerns about moral responsibility in machine ethics are obviously not unique to autonomous weapons. It is likely the case that if concerns about moral responsibility preclude our building robotic weapons, then those same concerns should preclude our automating more mundane functions as well. That is to say, we should be careful about treating autonomous weapons as too *sui generis*. The technologies that will enable more intelligent, more ethical machines seem poised to reshape our lives in remarkable ways, but the

impacts of such technologies, for better or worse, are largely opaque to us (Vallor). There is good reason to be concerned about the prospect of autonomous weapons, but we should also bear in mind the potential of these technologies to enable more discriminate, more humane outcomes in war, fully recognizing that such benefits are certainly not a foregone conclusion (Arkin). In either case, we should be concerned not to focus too narrowly on the technologies themselves, as opposed to the role they unavoidably have as a part of our human practices.

More to the point, it may not be realistic to think that we can opt out of lethal autonomous weapons. Others have pointed out the difficulty in enforcing a preemptive ban (Anderson and Waxman), but my concern lies more in the current state of technological progress. The technologies that will enable smarter and potentially moral machines are not unique to the military and they are not being driven primarily by military investments. Computer vision, like that being employed by Tesla's self-driving cars, has broad commercial applicability and will likely continue to see considerable private investment in research and development. This is likewise true for many other applications in robotics and artificial intelligence. Even if the international community were to adopt a ban on lethal autonomous weapons, the enabling technologies would continue to move towards making them a more realistic possibility. In this sense, an autonomous weapons ban is unlikely to stop weapons developers from incorporating enabling technologies into weapons designs. Instead, policy should focus on the manner in which human soldiers will exercise responsibility for these smarter and more capable weapons. Whether or not we can opt out of autonomous weapons, those soldiers cannot opt out of their moral responsibility for killing; good policy should start there.

Bibliography

Anderson, Kenneth, and Matthew C. Waxman. "Law and Ethics for Autonomous Weapon Systems: Why a Ban Won't Work and How the Laws of War Can." *SSRN Electronic Journal*, 2013. *Crossref*, https://doi.org/10.2139/ssrn.2250126.

Arkin, Ronald C. *Governing Lethal Behavior in Autonomous Robots*. CRC Press, 2009.

Asaro, Peter. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making." *International Review of the Red Cross*, vol. 94, no. 886, June 2012, pp. 687–709. *Crossref*, https://doi.org/10.1017/S1816383112000768.

DeLuca, Paul, et al. *Assessing Aegis Program Transition to an Open-Architecture Model*. Rand Corp., 2013, https://www.rand.org/pubs/research_reports/RR161.html.

Deputy Secretary of Defense. *Autonomy in Weapon Systems (DoD Directive 3000.09)*. Washington, D.C.: author, 21 Nov. 2012, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf?ver=2019-02-25-104306-377.

Eschleman, Andrew. "Moral Responsibility." *Stanford Encyclopedia of Philosophy*, Winter 2016, Metaphysics Research Lab, Stanford University, 2016, http://plato.stanford.edu/archives/win2016/entries/moral-responsibility/.

Finney, John W. "Navy Torpedo Mine Hunts Down Subs." *New York Times*, 15 Apr. 1974, p. 16.

Frederick, Jim. *Black Hearts: One Platoon's Descent into Madness in Iraq's Triangle of Death*. Harmony Books, 2010. *Open WorldCat*, http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=720365.

Gerdes, Anne. "Lethal Autonomous Weapon Systems and Responsibility Gaps." *Philosophy Study*, vol. 8, no. 5, May 2018. *Crossref*, https://doi.org/10.17265/2159-5313/2018.05.004.

Hammes, T. X. "Autonomous Weapons Are Coming, This Is How We Get Them Right." *The National Interest*, Dec. 2018, https://nationalinterest.org/blog/buzz/autonomous-weapons-are-coming-how-we-get-them-right-37532.

Law of War Manual. U.S. Department of Defense, June 2015.

McCaffrey, Barry. "Human Rights and the Commander." *Joint Forces Quarterly*, no. Autumn 1995, pp. 10–13.

McMahan, Jeff. "The Moral Responsibility of Volunteer Soldiers." *Boston Review*, Nov. 2013, http://bostonreview.net/forum/jeff-mcmahan-moral-responsibility-volunteer-soldiers.

"Mind the Gap: The Lack of Accountability for Killer Robots." *Human Rights Watch*, 9 Apr. 2015, https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots#_ftnref45.

Morris, Errol. *Fog of War: Eleven Lessons from the Life of Robert S. McNamara*. Sony Pictures, 2003.

Pfeifer, Jennifer H., and Sarah-Jayne Blakemore. "Adolescent Social Cognitive and Affective Neuroscience: Past, Present, and Future." *Social Cognitive and Affective Neuroscience*, vol. 7, no. 1, Jan. 2012, pp. 1–10. *Crossref*, https://doi.org/10.1093/scan/nsr099.

Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts. International Committee of the Red Cross, 8 June 1977, https://treaties.un.org/doc/Publication/UNTS/Volume%201125/volume-1125-I-17512-English.pdf.

Roff, Heather. "An Ontology of Autonomy and Autonomous Weapons Systems." *Unpublished*, 2019.

Roff, Heather M. "Killing in War." *Routledge Handbook of Ethics and War*, Routledge, 2013. *Crossref*, https://doi.org/10.4324/9780203107164.ch26.

Roff, Heather M., and David Danks. "'Trust but Verify': The Difficulty of Trusting Autonomous Weapons Systems." *Journal of Military Ethics*, vol. 17, no. 1, Jan. 2018, pp. 2–20. *Crossref*, https://doi.org/10.1080/15027570.2018.1481907.

Scientific Analysis Corporation, Paula, et al. *The Legal Status of Adolescents 1980*. U.S. Dept. of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation, 1981.

Shakespeare, William, and René Weis. *Henry IV, Part 2*. Oxford University Press, 2008.

Sharkey, Noel E. "The Evitability of Autonomous Robot Warfare." *International Review of the Red Cross*, vol. 94, no. 886, June 2012, pp. 787–99. *Crossref*, https://doi.org/10.1017/S1816383112000732.

Singer, P. W. *Wired for War: The Robotics Revolution and Conflict in the Twenty-First Century*. Penguin Books, 2010.

Sparrow, Robert. "Killer Robots." *Journal of Applied Philosophy*, vol. 24, no. 1, Feb. 2007, pp. 62–77. *Crossref*, https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Talbert, Matthew. "Moral Responsibility." *Stanford Encyclopedia of Philosophy*, Winter 2019, Metaphysics Research Lab, Stanford University, 2019, https://plato.stanford.edu/entries/moral-responsibility/#Psyc.

Taylor, Isaac. "Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex." *Journal of Applied Philosophy*, Nov. 2020. *Crossref*, https://doi.org/10.1111/japp.12469.

Vallor, Shannon. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press, 2016.

Walsh, Patrick. "The DOD Law of War Manual and Command Responsibility: Is It Time for a 'Necessary and Reasonable' Change to the UCMJ?" *Just Security*, 19 Aug. 2015, https://www.justsecurity.org/25488/dod-law-war-manual-command-responsibility-time-necessary-reasonable-change-ucmj/.