Philos. Technol. (2015) 28:125–137
DOI 10.1007/s13347-013-0138-3

SPECIAL ISSUE

Bridging the Responsibility Gap in Automated Warfare

Marc Champagne · Ryan Tonkens

Received: 10 January 2013 / Accepted: 11 October 2013 / Published online: 26 October 2013

Abstract Sparrow (J Appl Philos 24:62–77, 2007) argues that military robots capable of making their own decisions would be independent enough to allow us denial for their actions, yet too unlike us to be the targets of meaningful blame or praise—thereby fostering what Matthias (Ethics Inf Technol 6:175–183, 2004) has dubbed "the responsibility gap." We agree with Sparrow that someone must be held responsible for all actions taken in a military conflict. That said, we think Sparrow overlooks the possibility of what we term "blank check" responsibility: A person of sufficiently high standing could accept responsibility for the actions of autonomous robotic devices—even if that person could not be causally linked to those actions besides this prior agreement. The basic intuition behind our proposal is that humans can impute relations even when no other form of contact can be established. The missed alternative we want to highlight, then, would consist in an exchange: Social prestige in the occupation of a given office would come at the price of signing away part of one's freedoms to a contingent and unpredictable future guided by another (in this case, artificial) agency.

Keywords Robotics · Agency · Ethics · War · Responsibility

1 Introduction

If a robot capable of setting its own goals were to go on a killing spree, who would we blame? Or, if such a robot were to exercise its autonomy in a manner inconsistent with our understanding of morally permissible behavior, who could we justifiably hold responsible? The possibility of creating a robot with the ability to make

M. Champagne (✉)
Department of Philosophy, York University, 4th Floor, Ross Building South, 4700 Keele Street, Toronto M3J 1P3 Ontario, Canada
e-mail: gnosiology@hotmail.com

R. Tonkens
Centre for Human Bioethics, Monash University, E675 Menzies Building, Clayton VIC, Melbourne 3800, Australia
e-mail: ryan.tonkens@monash.edu

decisions on the basis of its own initiative(s) rather than preprogrammed commands, and of setting its own ends and/or achieving subordinate ends through means of its own devising, is admittedly a speculative idea at this stage. One can nevertheless imagine, as Sparrow (2007) does, a scenario where such military weaponry is somehow capable of making its own decisions without direct or indirect instructions from a human being. The robot may act in pursuit of an end dictated to it by another, but do so in ways that are unpredicted (unpredictable) and not in line with pursuing that end in a morally acceptable manner. Deploying these kinds of autonomous robots in the theater of war would have far-ranging ethical consequences because, in such contexts, not only are we dealing with life and death situations, we also expect the various parties to be held responsible for the choices that they make. Addressing a less speculative topic, Walzer (1977, p. 287) wrote that "[i]f there are recognizable war crimes, there must be recognizable criminals." It is natural to want to carry this principle over to automated warfare. However, since the hypothetical robots that interest Sparrow would be midway between us and plain machines, they would slip by this requirement of moral responsibility: They would be independent enough to allow humans plausible denial for their actions—yet too unlike us to be the targets of meaningful blame or praise.

Such autonomous yet unfeeling robots would therefore inhabit a morally ambiguous space Matthias (2004) has aptly dubbed "the responsibility gap." Any unethical act on their part would therefore betoken the most senseless event conceivable, a moral failure for which no one could be fairly held responsible. Understandably, then, Sparrow thinks this would be enough to bar the use of autonomous "killer robots" in war.

We follow Sparrow (2007) in believing that one condition for being an agent that can be held morally responsible is the capacity to be punished/rewarded through proportionate suffering/remuneration (pp. 71–73). Although other conditions exist (e.g., willful intent), falling within the ambit of punishment and reward is arguably a necessary condition of moral responsibility. Thus, insofar as autonomous robots lacking sentience could not be punished in any meaningful way (Sparrow 2007, p. 73), they would not qualify as responsible agents. However, since these machines could autonomously choose their ends, they would be culpable, and since they act deliberately, they would also be liable. Hence, such robots should be held responsible for their actions (if possible), and we owe it to (at minimum) the parties in the contexts where the robots would be in use to allocate responsibility and seek proportionate (yet fair) retribution.

We thus agree with Sparrow that, following the received tenets of just war theory and international laws of war, one or more individuals must be held responsible for all actions taken in a military conflict. Yet, when Sparrow concludes from this that autonomous robots should not be deployed, we believe his argument proceeds a bit too hastily. Indeed, we want to suggest that there is a way to seal the responsibility vacuum Sparrow points to. Although we do not endorse the use of such military devices, we think Sparrow's argument overlooks the possibility of "blank check" responsibility: A person (or persons) of sufficiently high military or political standing could accept responsibility for the actions (normal or abnormal) of all autonomous robotic devices—*even if that person could not be causally linked to those actions besides this prior agreement*.

The basic intuition behind our proposal is that two things can be related just in virtue of their being related. That is, we humans retain the liberty to *impute* relations even when no other form of contact can be established. This is what humans do, for example, whenever they link a symbol to a referent (e.g., the word "chair" standing for worldly chairs). The relation at play in such instances is mind-dependent, in that it would not hold were it not for an agreement between agents. Unlike causal links, one cannot discover an evidential basis for such imputed relations, since they are made, not found. In an ethical context, noncausal imputation is akin to what is routinely called scapegoating (i.e., deliberately singling out a particular person as the recipient of responsibility or blame). But, this possibility should not overshadow the fact that, *when informed consent is present*, the same mechanism can be used fairly. Indeed, we argue that consent can secure the sought-after component of responsibility through a person's very willingness to partake in a contractual agreement. As will be explained below, by willingly agreeing to the terms of a contract, the informed agent(s) *impute*(*s*) *responsibility on herself* for the actions of an autonomous machine. The missed alternative we want to highlight, then, would essentially consist in an exchange: Social prestige in the occupation of a given office could come at the price of signing away part of one's freedoms to a contingent and unpredictable future guided by another (in this case, artificial) agency.

We shall not be arguing that holding the office in question would be conditional on one's willingness to actually use killer robots. Our more prosaic point is that *if* such robots are deployed, then someone (or several people) can willingly and publicly accept responsibility for whatever ensues. Obviously, this proposal leaves open the question of whether the antecedent of this conditional is or should be affirmed. Still, to the extent that such agreements can be fairly and reasonably implemented, Sparrow's case needs tinkering.

We will present our argument in a fairly straightforward manner. We will first describe and motivate the problem, then we will articulate and defend our solution.

2 Sparrow's Dilemma

Sparrow recognizes the common criticism (Asaro 2008; Krishnan 2009; Singer 2010) that the advent of military robots may trivialize entry into war. His argument against the use of autonomous robots, though, is more subtle and turns on a disturbing prospect of across-the-board moral blamelessness. Even if we grant the (debatable) cognitive scientific prognostication that it is possible for a nonsentient robot to be autonomous—in the sense not just of being able to determine which means best suit a given end, but in the stronger sense of actually determining the best ends to follow—we are left with an ethical dilemma. Specifically, the question arises whether there is any room left for the attribution of blame (and praise) where these robots are deployed.

As Sparrow points out, if we are ever to conduct wars utilizing such technology and simultaneously maintain our standard ethical inclinations regarding responsibility for one's actions, we must ask ourselves who would be responsible in the event of a violation by an autonomous robot in the conduct of war. Sparrow's claim, in short, is that no one could be rightly held responsible for any actions (atrocities or otherwise) perchance committed by this type of robot. Since such a situation of blamelessness

would be morally unacceptable (and unjust), prohibition of their use in war seems to follow.

So stated, this problem is a general one, and in principle arises anytime (a) some autonomous entity is responsible for its behavior yet (b) holding that entity responsible would be absurd. It just so happens that, with respect to adult human military personnel, those that are (and need to be) responsible for their actions can all be held responsible, as they meet the conditions for being the targets of meaningful praise or blame (among other things, perhaps). Although an autonomous robot should likewise be responsible for its actions, it cannot be held as such.

To make this worry vivid, consider a fully self-determining military machine equipped to identify the intention(s) of combatants. Imagine a scenario in which such a robot would lethally engage an enemy platoon that was clearly surrendering before and during its attack (Sparrow 2007, p. 66). Such a situation would leave us with the following tension. To the extent that an "autonomous robot" fits its twofold label, humans cannot be blamed for its actions, since the robot has genuine "autonomy." Yet, the robot itself cannot be blamed for its actions either, since it is merely "robotic" and thus impervious/oblivious to any meaningful form of punishment (A sentient grasp of events is pivotal and explains why, say, someone being whipped is given a mouth guard and not a sedative or anesthetic). The robot would presumably have nothing of moral worth at stake, like freedom that can be taken away through imprisonment. If we want to keep moral considerations in the picture and retain the idea that just war theory and the international laws of war require us to justifiably hold someone responsible when things go wrong, then the situation described by Sparrow becomes very problematic.

It is important to underscore that the moral predicament discussed by Sparrow arises only when the violent act of the robot is sandwiched between a narrow set of circumstances. Specifically, the robot must be sophisticated enough to make its own choices—but not so sophisticated that it can experience pain and pleasure. It is thus worth stressing that the range of robots Sparrow's argument applies to is very restricted. All of the semiautonomous or remote military robots currently in existence are irrelevant to his discussion, since in the use of these robots humans can at all times be held responsible, say, via their contribution to a program's content or a device's activation. Sparrow (2007), by contrast, wants to call our attention to a far more difficult prospect. Hence, it is crucial to be clear on what exactly is meant by a fully "autonomous" robot in this context:

> Artificially intelligent weapon systems will thus be capable of making their own decisions, for instance, about their target, or their approach to their target, and of doing so in an 'intelligent' fashion. While they will be programmed to make decisions according to certain rules, in important circumstances their actions will *not* be predictable. However, this is not to say that they will be random either. Mere randomness provides no support for a claim to autonomy. Instead the actions of these machines will be based on reasons, but these reasons will be responsive to the internal states—'desires', 'beliefs' and 'values'—of the system itself. Moreover, these systems will have significant capacity to form and revise these beliefs themselves. They will even have the ability to learn from experience. (p. 65)

The robots Sparrow envisions would thus have the autonomy to reject their programmed rules or to apply those rules in ways that are not in line with judgments we would endorse. To the extent a robot would do this, it would think outside *our* box.

No matter how remote it may be, Sparrow wants to address this possibility directly, before it arises (a laudable and prudent preemptive attitude which, by engaging in earnest with the subject matter, we obviously share). Sparrow is thus more concerned with guarding against a responsibility vacuum than with arguing that limitations should be placed on the autonomy of machines. Placing an inviolable limit inside a robot's mind is, on this view, no more satisfactory than placing a robot's body inside an inescapable box: Both strategies get the job done, but only by avoiding the philosophical challenge. The moment such artificial constraints are placed on a military robot's range of conduct, be it at the level of hardware or software, that robot no longer falls within the ethically problematic range Sparrow is interested in.

As an example of this, consider Arkin's (2009) suggestion that autonomous lethal robotic systems be equipped with an in-built design component providing us with clear data regarding who is responsible for the robot's actions. This could include explanations for using or omitting lethal force prior to deployment in a specific mission context. Such "a responsibility advisor" (as Arkin calls it) "makes explicit to the robot's commander the responsibilities and choices she is confronted with when deploying autonomous systems capable of lethality" (2009, p. 145). Mock guilt could even be made into a quantifiable variable influencing a robot's behavior in a given military conflict. For instance, if a guilt value associated with a given action exceeds a certain threshold, the robot could have a mechanism that reprimands it for that action (Arkin and Ulam 2009). Although such a buffer is clearly meant to remedy the violence robots may cause, a proposal like this does not truly address Sparrow's dilemma, insofar as reinstating straightforward human responsibility by shaving off a robot's leeway in the election of its ends is tantamount to dodging the issue at hand. Irrespective of its practical merits then, Arkin's proposal ultimately misses the point.

This inadvertent tendency to alter the scenario into something more manageable is ubiquitous. Petersen (2012), for example, advocates robot servanthood. Similarly, Lin et al. (2008, p. 54) argue that one can guarantee responsibility for war crimes by mandating a "slave morality" for autonomous military robots. Applied ethics merely "applies" ethical standards, so rehabilitating a Nietzschean two-tiered morality to expediently resolve difficult cases still constitutes a philosophical endorsement of that double standard. Alas, "competing moral claims generated by different moral frameworks does not show much beyond the fact that *different moral frameworks may generate competing moral claims*" (Tonkens 2012, p. 140; emphasis in original). To be principled, an appeal to slave morality cannot be *ad hoc* or arbitrarily confined only to problematic instances. It is unclear, though, whether proposals that invoke this double standard fully appreciate the ramifications of allowing the conjunction of autonomy and slavery. All other things being equal, if a less onerous way to address the difficulties at hand can be found, it should be preferred.

The suggestion of Lin and colleagues to instill a slave morality in military robots may, if implemented effectively and univocally, help to *avoid* Sparrow's dilemma, but it does not *solve* it. As stressed earlier, we are following Sparrow in considering the

case of military robots that have a very high level of autonomy, whereas Lin et al. (2008, p. 64) deal with robots that are not so autonomous as to be choosers of their ends. The former situation is closer to the human condition, insofar as adult rational animals have a say when they adopt/endorse a moral code. Thus, the robot we (and Sparrow) are considering is one that can genuinely disobey its orders in a manner not consistent with the ethics of warfare. Clearly, should one ever be faced with robots that are autonomous in just this sense, arguing that "They should have been programmed with a slave morality in the first place" would be unsatisfactory. This would be no better than saying "If only serial killers were kittens, the problem would be solved." Slave morality is no solution or a solution in a different topic.

Of course, there is a disanalogy in the sense that, unlike transforming into a kitten, conduct can change. Yet, if we are committed to keeping the terms of the debate intact, the only aid this could bring would be if robots were to freely elect to be submissive toward humans (if they had no choice in the matter, they would not be the sort of robots we are interested in). In volunteering themselves into bondage, each robot would have to make (and sustain) this decision individually and do so on the basis of deliberations that cannot be mechanically binding. Any fortuitous turn of events that benefits humans still benefits humans. As far as solutions go, though, one can only *hope* for this to occur.

Taking another route, Lokhorst and van den Hoven (2012, p. 149) have recently challenged Sparrow's conclusion by arguing that there may be effective alternatives available for remedying unacceptable behavior, namely by adding sentience to the mix. Their proposal, however, stems from a misunderstanding of Sparrow's concerns. Sparrow is willing to (agnostically) grant Lokhorst and van den Hoven's claim that advancements in AI and robotics research could conceivably invest autonomous robots with the ability to suffer. Yet, these are not the sorts of robots that trouble him, since here we could presumably hold something responsible, namely the robot itself. Indeed, to the extent robots could suffer in the relevant way, they would become legitimate targets for genuine punishment, and so the quandary of moral blamelessness previously canvassed would not appear. Likewise, Lokhorst and van den Hoven's proposal to adjust a violent robot's programming so that it does not commit a similar atrocity in the future does not satisfy the demand that someone be held responsible for the previous misbehavior. Even if we could successfully tweak a robot's programming after we detect a flaw, the initial act that generated our *ex post facto* intervention would remain unpunished. To be sure, preventing similar atrocities from occurring again is a very important task. Yet, this still leaves us in a responsibility vacuum with respect to the original act—a situation whose resolution is *also* very important.

With these qualifications in mind, Sparrow contends that, if a robotic agent is truly acting autonomously, then we are not justified in holding anyone responsible for the actions of that agent. To be sure, if a robot kills innocent civilians following its own internal states (like beliefs, desires, and goals), an understandable reflex will be to blame that robot. Autonomous war machines, however, could not be blamed because, without a real capacity to suffer, they would lack a characteristic required to be the subject of meaningful punishment. Sparrow thus argues that, given the appropriate level of machine autonomy, any transgressions made by a robot give rise to an ethical catch-22. Since there are no suitable candidates for satisfying the *jus in bello* clause,

all the available possibilities seem either irrational or unfair: If we blame a nonsentient machine, we are being irrational, and if we blame someone unrelated to the act, we are being unfair. Given that these two premises are acceptable, their troublesome conjunction demands our immediate philosophic attention.

Let us look at the first horn of this dilemma. Although they can certainly be destroyed or damaged, robots do not have anything of comparable moral worth at stake in the outcome of their actions, no welfare that could be compromised, no personal freedom that could be truncated. As Sparrow (2007) notes,

> In order for any of these acts to serve as punishment they must evoke the right sort of response in their object […]. I shall assume that the most plausible accounts of the nature and justification of punishment require that those who are punished, or contemplate punishment, should suffer as a result. While we can imagine doing the things described above [i.e. fines in earnings, imprisonment, corporal or capital punishment] to a machine, it is hard to imagine it suffering as a result. (p. 72)

This lack of an experiential dimension, Sparrow argues, essentially makes it futile to hold the machine responsible for its transgressions.

It could perhaps be objected that courts and lay folk hold corporations responsible in certain contexts, even if corporations do not meet Sparrow's criteria for being punishable. Courts and lay folk could of course be misguided when they stretch the semantics of "punishment" in this (animistic) direction. Even so, the objection would be that while corporations are not sentient, punishing them is not futile. However, this objection would have to overlook the fact that "punishing" a corporation through fines and/or restrictions only works because it negatively affects the interests of its (sentient) constituents, say, by lowering the financial gain of certain humans (who would have stood to enjoy those gains). Neither the parts nor the wholes of autonomous robots can suffer in this way. Imprisonment severely restricts the range of projects one can undertake, but such a set of descriptive circumstances does not have any normative significance unless the organism at the center of those circumstances yearns to see its projects fulfilled (see Champagne 2011). Truncating a robot's autonomy by punishment would only be successful if a robot had a genuine interest in being free and the emotional and psychological capacities to appreciate the value of such freedom. With the absence of these dimensions, the dread of being incarcerated is not liable to move the sort of robot under discussion.

The other horn of the dilemma is more straightforward. Clearly, it is wrong to blame someone for actions she did not partake in or endorse. Once we grant that all robots and people are inadmissible, a tension naturally follows. Indeed, just war theory demands that someone be responsible for military actions (see for instance the customs of war described by the International Committee of the Red Cross). Since blaming someone at random would be wrong, developing and using autonomous military robots is morally unacceptable.

This, at any rate, is what Sparrow argues. What we want to suggest now is that, when carefully used, consensual imputation effectively bridges the responsibility gap that concerns Sparrow.

3 A Way Out of Sparrow's Dilemma

Here, in programmatic outline, is how we think the dilemma canvassed by Sparrow can be overcome. Should fully autonomous robotic agents be available to a military, the decision to deploy these robots could fall on the highest person in command, say, the president (or general, etc.). The president would be fully aware that the autonomous robots are capable of making their own decisions. Yet, a nonnegotiable condition for accepting the position of authority would be to accept blame (or praise) for whatever robotic acts transpire in war.

Admittedly, the theater of war is foggy. Yet, with the rigid imputation of blame secured, if the deployment of these machines renders the resulting war "foggy" to a degree which makes the authority figure uncomfortable with accepting her surrogate responsibility for the robots' actions, then it follows that these robots should not be deployed. On the assumption that it would be in a democratically elected official's best interest to wage an ethical war (if it is ethical to wage a war at all), he or she would likely ensure that these robots will abide by accepted international rules; otherwise, she would have excellent reason *not* to allow those robots to be used. Yet, the force of our proposal lies in the fact that the requirement to accept "blank check" responsibility would hold even if the autonomy of the robot was shown to be unconstrained. To be sure, if the president or military general would be unsure about whether these robots would indeed behave morally, then prudence would dictate that she not consent to their deployment in the first place. There may be some ethical issues that emerge here with respect to providing this option to sufficiently ranking individuals in practice. Yet, the decision of whether or not to sign such a contract, once presented, is not necessarily a "problematic option," that is, a choice one would have been better off never having had to make (Velleman 1992). And, insofar as the contract should only be entered into willingly (through the exercise of informed, unforced, consent), then there are safeguards in place to prevent morally dubious ways of soliciting or recruiting individuals to sign such a contract, thereby accepting responsibility for the killer robots. If the individual is not willing, then they should not agree to the terms of the contract; if they are unwilling to agree and yet are forced to accept the contract, then they would not be liable, and arguably their coercer would be responsible for the behavior of the autonomous machines. Be that as it may, the moral cost of any gamble would fall squarely on the gambler, who would be readily identifiable for all to see. In this way, we could at all times ensure that a human is liable to receive blame for any self-initiated robot acts deemed immoral.

This satisfies the aforementioned requirement that we justly hold *someone* responsible for the actions taken in war. Sparrow (2007), by contrast, argues that it is unfair to "assign" responsibility to the commanding officer of autonomous robots, since "[t]he use of autonomous weapons therefore involves a risk that military personnel will be held responsible for the actions of machines whose decisions they did not control" (p. 71). On our account, Sparrow's worry can be accommodated such that the "risk" he highlights is eliminated. Our picture could be complicated further by the inclusion of administrative safeguards so that it is formally expressed by the commander (in addition to her superiors) that she is (un)willingly following orders to deploy such machines and accepts (no) responsibility for the outcome of doing so. In this way, even if the topmost official does sign off on the use of such machines, the

frontline soldiers using them could formally express their (un)willingness to do so and hence their (un)willingness to accept partial responsibility for its actions. After all, blank checks are indeterminate in the amount they allow, but crystal clear about who is doing the allowing.

Sparrow (2007) briefly considers (and dismisses) something along the lines of what we are proposing. He writes that "[i]n these cases we *simply insist* that those who use [the robotic weapons] should be held responsible for the deaths they cause, even where these were not intended" (p. 70; emphasis added). The main difference between our proposal and the one discussed by Sparrow is that we regard it as fair and reasonable for the commanding officer to willingly and freely *assign responsibility to herself*, ahead of time. It is standard to use informed consent as part of our gauge of the ethics of intersubjective practices (e.g., patient recruitment for experimental research), and with such consent in place, (at least some) responsibility is thereby transferred to the agent doing the consenting.

It is important to underscore that our way out of Sparrow's dilemma does not entail that a community will arbitrarily select a prominent figure as a lightning rod for its disapproval. Rather, we think the incentives are so balanced that users of autonomous machines could willingly accept responsibility for their actions, ahead of time, and thus become a justified, reasonable, and fair locus of surrogate responsibility for those autonomous robots' actions. We still cannot hold a robot responsible for its actions (and in a real sense we still should hold the robot responsible), and merely assigning the commanding officer responsibility is still unfair. Yet, asking a suitably ranked and sufficiently informed person (or persons) to decide whether or not she is willing to prospectively assume responsibility for the actions of autonomous robots *is* fair and would satisfy the requirement that someone be held responsible. Perhaps there are other, unstated, desiderata that we might want to take into consideration, but as things currently stand, the proposal goes through.

It is therefore appropriate to understand our view as a sort of "vouching for" rather than a "pointing of fingers." Whether this practice will be well implemented is an empirical issue that cannot be determined in advance of the facts. But, we can nevertheless infer that, if an informed surrogate willingly consents, she can take the risk identified by Sparrow (there may of course be independent reasons for commanding officers or presidents to *not* accept surrogate moral responsibility).

Although Sparrow does not seriously consider the possibility of establishing a viable social contract of the sort we gesture at, the basic idea has proven its merit in the customary chain of command of the military, where there is already divvying of responsibility (McMahan 2009). Our proposal is not a novel one in this sense. For instance, when the captain of a ship accepts blame for the actions of those officers under her, the terms of that office permit us to blame the captain, even if no clear causal chain can be established that would link her to the reprehensible action(s) in question. Accepting responsibility for such behavior is simply part of the captain's job description, a "role responsibility" to which the captain has committed through her explicit acceptance of her post. For example, the Canadian Armed Forces (via the 1985 *National Defence Act* and the Department of National Defence's *Army Ethics Programme*) subscribes to this way of assigning responsibility for ethical and legal misconduct. A similar rationale for the assignment of responsibility was at work in the Nuremberg trials, where commanding officers were deemed (partially) responsible for the behavior

of their troops (although, in some cases, the troops were also held responsible). The same sort of contractual acceptance of responsibility could be fruitfully employed, we argue, in the case of autonomous robots.

Sparrow's (2007) analogy between autonomous killer robots and child soldiers (pp. 73–74) can thus be granted without compromising the force of our critique. Child soldiers are not such that they lack autonomy (in its entirety), and yet they do not seem responsible for their actions—blaming them for their actions seems unjustified, even if not entirely unwarranted. Parents are not typically held responsible for the actions of their grown offspring. Still, given that parents *are* responsible for the actions of their children (who are arguably autonomous in the required sense), it seems fair to say that part of the role of parent is to own up to this responsibility, rearing the child so that they learn to behave in (morally) acceptable ways, and accepting (partial) responsibility in cases where she intentionally fails to do so. Needless to say, the addition of an explicit social contract would make that link even tighter.

It is worth bearing in mind that our proposal charitably grants Sparrow's twin contentions that (1) a programmer cannot be justifiably held responsible for the actions of a truly autonomous robot and that (2) holding a nonsentient robot responsible is meaningless and unsatisfactory. What we are challenging is the assumption that this exhausts the moral avenues. In an effort to introduce a *tertium quid*, we argue that if a commanding officer willingly deploys autonomous robots that can act immorally and moreover publicly recognizes that those robots cannot be held responsible, then she has thereby accepted responsibility for their actions. A person in this position is exercising her privilege of uncoerced rational consent. This in turn yields a clear—but noncausal—connection between the officer and the actions of the autonomous robot, insofar as the person who directly sanctions the use and deployment of such robots assumes responsibility for their actions come what may and is justifiably subject to punishment for doing so.

We need to ensure that autonomous robots meet ethical and safety standards prior to their deployment. It seems reasonable to assume that a commanding officer faced with the decision to deploy autonomous robots would work hard to ensure that those robots behave properly, since it is *she* who would be held responsible and punished in the event of their misbehavior (however one wants to construe this). Indeed, it would make for an informative empirical study to ask current military officers and commanders ahead of time whether they would be willing to accept such surrogate responsibility. Coupled with our proposal, maybe the results of such an inquiry could suffice to halt the development of autonomous war machines. After all, why build autonomous yet unpunishable lethal machines if no one is willing to accept responsibility for their actions? Although long-range missiles sometimes cause collateral damage (Sullins 2010, p. 268), we do not blame the missile itself; instead, we blame those deploying the missile (Asaro 2008, p. 51). The case of US military practice actually simplifies this by requiring all equipment in the field to be certified fit for service so that the operator, not the manufacturer, has sole responsibility. Autonomy, however, introduces a considerable element of risk in deploying high-tech weaponry, since neither the manufacturer nor the "operator" (the term now becomes a misnomer) can be fully certain of the decisions that a given robot may take. By issuing a blank check, our noncausal tether is loose, in the sense that it

makes provisions for uncertainty; yet, it is tight, in the sense that, whatever happens, blame/praise will be bestowed at the human end of the relation. This is a tradeoff for the privileges one enjoys from being a commander. If, as a contingent matter, no one happens to submit to this demand, then so much the worse for those who want to create and deploy autonomous robots.

A cynic could argue that, once proper punishment has been carried out, a fresh candidate could come into the picture and allow the carnage to start all over again. Of course, so long as forecasts are being had on the cheap (being answerable to nothing more tangible than one's imagination), nothing bars these kinds of criticisms. Realistically though, humans are responsive to past events, so (all other things being equal) the inductive likelihood that a person in a visible position of authority would sign off after such repeated debacles should shrink. Again, for someone who inclines to worst case scenarios, that may bring little comfort. Yet, what is the alternative: Rogue use of killer robots without any attempt at securing a morally responsible agent? Taking the cynic at her word, she should recognize that, either way, technologies that can be used will be used. Once that maxim has been granted, the task becomes to minimize undesired outcomes. At any rate, it is inconsistent for one to claim that (1) repeated episodes of carnage will naturally ensue, and (2) if we just say no, everyone will abide by that. (1) betokens unwarranted pessimism, (2) is naively optimistic—and oscillating between the two would be *ad hoc*.

Our suggestion that the likelihood of deployment would shrink in light of past fiascoes of course assumes that the relevant practical decisions would be motivated by a certain measure of informed rationality. This assumption, like all assumptions, can certainly be called into question. However, the retort that humans have the power to act contrary to the dictates of reason is true but trivial, since the observation applies to the full spectrum of philosophical branches concerned with human activity. The burden is thus on whoever wants to exploit the element of voluntarism inherent in decision making (Ullmann-Margalit and Morgenbesser 1977) to show why glass-half-empty worries about human wickedness should be deemed more likely than glass-half-full confidence in human reasonableness.

In the same vein, we might note that, despite the focus of Sparrow's discussion, full autonomy does not solely imply a willingness to kill. Presumably, such an exotic technological endowment, if it is indeed possible, might also lead an autonomous robot to mimic laudable human acts. As such, the “blank check” we have introduced need not be viewed solely in a negative light. For the exchange to fully succeed, it is not enough to limit the ethical responsibility to a consenting human commander when things go wrong—we could/should also transfer any praise a robot might garner to the same person. While it is perhaps less intuitive to connect the success(es) of an autonomous robot with a human who had very little or nothing to do with the mission in question, allowing for such a noncausal transfer of prestige could be a vital component in the social contract we envision.

4 Conclusion

The hypothetical situation presented by Sparrow points to an interesting set of circumstances which force us to question our notion of ethical responsibility in

increasingly unmanned battlefields. In that sense, it betokens philosophical reflection on technology at its best. The twofold recognition that neither hapless programmers nor unfeeling machines deserve punishment acts like a dialectic vise, thereby compelling Sparrow to conclude that the use of autonomous robots would be a moral aberration. There is a sense in which he is clearly right. We have argued, however, that Sparrow's analysis, while thought provoking and in many ways correct, nevertheless overlooks the possibility of a sufficiently high-ranking commanding officer (or officers) accepting responsibility for the robot's actions, and thus being accountable for any violation of the rules for the ethical conduct of warfare.

In essence, our proposal retains the noncausal imputation involved in scapegoating while dropping its arbitrariness: Since humans are capable of informed consent and pleasure/pain, a suitable and ascertainable target for punishment can be established, thereby ensuring visible conformity with the tenets of just war theory. While victims of the immoral behavior of autonomous lethal robots may not always be satisfied with the level of retribution gained (e.g., the forfeiture of office/rank, fines, imprisonment), it is important that punishment for such actions go only to those that deserve it, and getting some fair retribution is better than not getting any at all.

Although it may not be possible to trace a tenable physical link between the negative (or positive) actions of a fully autonomous robot and its human commander, our suggestion has been that this transitive chain is inessential to the ethical question and can be bypassed by using a more explicit social contract or "blank check." A robot may perform self-initiated actions, but it does not have to suffer what we would consider a just punishment for a violation of our ethical rules. Instead, the moral blame and accompanying punishment could be placed squarely on a human agent (or agents) who, through her own volition, has traded a part of her freedoms for the prestige of occupying a high-ranking position in a given social hierarchy (say, a governmental or military chain of command). If no one is willing to accept this responsibility, then they should not deploy autonomous killer robots in the first place. In either case, this way of framing the issue keeps a human in the loop in a morally defensible manner, rendering the force of Sparrow's argument weaker than it first seemed.

One could object that having a human take moral responsibility would indeed offer a good compromise for policy-making purposes, but that this would fall short of an ethically acceptable way to link moral reasonability of a sentient agent to the actions performed by a robot. Such an objection rests on the enthymeme that causal authorship of (or some weaker mind-independent link to) an action must be present. Our suggestion is that it is not mandatory to accept this enthymeme and that automated warfare presents one situation where it might wisely be jettisoned.

Our critique does not contradict Sparrow's case for peace, but instead shows that that case is incomplete. We have thus addressed the overlooked option right away. Speculation must eventually pass through the bottleneck of action, so when all is said and done, Sparrow's prescriptive yield is "Wait, don't push that button, it might lead to senseless violence." Our yield is "Wait, don't push that button, it might lead to senseless violence, and if it does, you will be held responsible and punished."

Acknowledgments We would like to thank Mohammad Al-Hakim, Robert Myers, Joanna Bryson, David Gunkel, Luciano Floridi, and Alberto Richards, participants at the 2012 joint meeting of the International Association of Computing and Philosophy and the Society for the Study of Artificial Intelligence and the Simulation of Behaviour, as well as four anonymous reviewers for this journal. We would also like to acknowledge the support of York University's Department of Philosophy.

References

Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. Boca Raton: Chapman and Hall.

Arkin, R.C., Ulam, P. (2009). An ethical adaptor: behavioral modification derived from moral emotions. Technical Report GIT-GVU-09-04, Georgia Institute of Technology, Atlanta

Asaro, P. M. (2008). How just could a robot war be? In A. Briggle, K. Waelbers, & P. Brey (Eds.), *Current issues in computing and philosophy* (pp. 50–64). Amsterdam: Ios Press.

Champagne, M. (2011). Axiomatizing umwelt normativity. *Sign Systems Studies, 39*(1), 9–59.

Krishnan, A. (2009). *Killer robots: legality and ethicality of autonomous weapons*. Farnham: Ashgate.

Lin, P., Bekey, G.A., Abney, K. (2008). Autonomous military robotics: risk, ethics, and design. Report for the US Department of Defense/Office of Naval Research

Lokhorst, G.-J., & van den Hoven, J. (2012). Responsibility for military robots. In P. Lin, G. A. Bekey, & K. Abney (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 145–156). Cambridge: MIT Press.

Matthias, A. (2004). The responsibility gap. *Ethics and Information Technology, 6*(3), 175–183.

McMahan, J. (2009). *Killing in war*. Oxford: Oxford University Press.

Petersen, S. (2012). Designing people to serve. In P. Lin, G. A. Bekey, & K. Abney (Eds.), *Robot ethics: the ethical and social implications of robotics* (pp. 283–298). Cambridge: MIT Press.

Singer, P. W. (2010). *Wired for war: the robotics revolution and conflict in the 21st century*. New York: Penguin.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Sullins, J. P. (2010). RoboWarfare: can robots be more ethical than humans on the battlefield? *Ethics and Information Technology, 12*(3), 263–275.

Tonkens, R. (2012). Out of character: on the creation of virtuous machines. *Ethics and Information Technology, 14*(2), 137–149.

Ullmann-Margalit, E., & Morgenbesser, S. (1977). Picking and choosing. *Social Research, 44*(4), 757–785.

Velleman, J. D. (1992). Against the right to die. *The Journal of Medicine and Philosophy, 17*(6), 665–681.

Walzer, M. (1977). *Just and unjust wars: a moral argument with historical illustrations*. New York: Basic.