# Punishing Robots – A Way Out of Sparrow's Responsibility Attribution Problem

Abstract: The Laws of Armed Conflict require that war crimes be attributed to individuals responsible and punished. Yet assigning responsibility for the actions of Lethal Autonomous Weapon Systems (LAWS) is problematic. Robert Sparrow argues that if specific agents cannot be fairly and reasonably held responsible for war crimes committed by such systems, then LAWS lack legal and moral legitimacy. He further argues that neither the programmers and engineers creating truly autonomous systems, nor their commanders, nor the machines themselves can be held responsible for the actions of LAWS. This would be unfair in the case of the humans and impossible in the case of the machines, which cannot be punished as they lack the capacity for phenomenal experience. I challenge the latter claim by showing that all the morally desirable goals that punishment aims for in humans – incapacitation, rehabilitation and deterrence – can be effected in robots by alternative but more reliable means. My account focuses on describing how the behaviors enforced by deterrence in humans may be achieved via a mixture of prevention and threat of goal frustration. The only aspect of punishment that cannot be replicated in LAWS is the retributive one. However, incapacity for suffering retribution would delegitimize LAWS only if significant moral value was to be attributed to purely retributive punishment, and if no other considerations could outweigh retribution's value. Since these other reasons include the need to spare combatants from death and injury, such a position is untenable.

One of the many issues within the debate over legal and moral feasibility of Lethal Autonomous Weapon Systems (LAWS) is the problem of responsibility attribution. Presently, all applications of lethal force in warfare may in principle be traced to the agent responsible. In cases when lethal use of force is deemed illegitimate this agent may be held accountable for a war crime. There are many valid and uncontested reasons for maintaining such accountability. Failure to hold war criminals accountable is itself a breach of the Laws of the Armed Conflict (LOAC). It follows that if a LAWS is introduced into combat, some agent needs to be held responsible for its actions, especially ones involving the use of lethal force. It is not immediately clear who this agent should be. Not any person will do, as it has to be someone who has sufficient influence on the behavior of the robot to be able to restrain it motivated by fear of legal and moral sanction. It would be not only unfair and unjust, unreasonable and ineffective to assign responsibility to someone without such influence.

Robert Sparrow (2007) has prominently argued that there is no sufficiently good solution to this problem, and that therefore the deployment of LAWS would be an illegal and immoral act. I will challenge this conclusion and certain presuppositions behind it. Most importantly, I will argue that all the functions of responsibility attribution and resultant punishment can be achieved in LAWS by other means, the retributive function of punishment being the exception. After explaining how other functions of punishment can be achieved in non-sentient systems, I claim that incapacity to suffer retribution cannot on its own be regarded to outweigh all other moral reasons for allowing LAWS on the battlefield. At the very least, it provides a much weaker justification than inability to be deterred from certain behaviors, posited by Sparrow.

Premise I of Sparrow's argument states that allowing unattributed use of lethal force on the battlefield is criminal and immoral. Premise II is that fair and meaningful attribution of responsibility is not possible for LAWS. Sparrow identifies three classes of agents who could be made liable for the actions of LAWS – programmers, commanders and the machines themselves. He argues that programmers and commanders do not have sufficient control over the actions of LAWS to be held responsible, and the machines themselves are not capable of receiving punishment and thus answering for their crimes or being deterred from commission thereof. Regarding programmers, Sparrow states:

> "The connection between the programmers/designers and the results of the system, which would ground the attribution of responsibility, is broken by the autonomy of the system. To hold the programmers responsible (…) would be analogous to holding parents responsible for the actions of their children once they have left their care." (2007, 70).

Sparrow also denies that commanders may be held responsible:

> "(…) the autonomy of the machine implies that its orders do not determine (although they obviously influence) its actions. The 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. The more autonomous the systems are, the larger this risk looms. At some point, then, it will no longer be fair to hold the Commanding Officer responsible for the actions of the machine." (2007, 71)

I do not contest any of these assertions – they seem undeniable for genuinely autonomous systems. However, Sparrow's discussion of the third possibility – the LAWS themselves being held responsible – is very problematic. Sparrow proposes that a robot cannot be punished, unless robots become sentient – able to suffer – and thus moral patients as worthy of protection as the soldiers they are to replace, thus defeating the moral purpose of the whole enterprise[1] (2007, 73). The philosophically interesting subject of robot sentience aside, this presupposes that either 1) the means of disciplining robots are restricted to the types of means that can be used to discipline humans, or that 2) the retributive aspect of punishment is of critical importance to LOAC and ethics of war. Both of these suppositions are unfounded.

It is clear from Sparrow's comments about capacity for suffering punishment that he understands "holding responsible" to mean "suffer a legal sanction". I do not believe he is mistaken in equating being held responsible with being liable to punishment, as in the context of wartime crimes this assertion makes perfect sense. Sparrow's error is to insist that the

[1] Sparing sentient combatants death and suffering is not the only conceivable purpose of introducing LAWS – greater precision and ability for self-sacrifice resulting in lower civilian casualties, better civilian control over the armed forces or automating collective defense pledges, thus increasing their credibility and better deterring potential conflicts, would all constitute morally worthy reasons for developing them.

retributive aspect of punishment has to be preserved either because of the utility it generates and that cannot be achieved through a different mechanism, or because of its inherent value.

There are four commonly distinguished functions of punishment – deterrence, retribution, rehabilitation and incapacitation (Demlaitner 2015, 942). Of these four, rehabilitation and incapacitation are uncontroversially applicable to non-sentient LAWS. To incapacitate a LAWS is merely to switch it off or destroy it, whereas to rehabilitate it is to reprogram and/or retrain it[2]. As by definition incapable of suffering, non-sentient LAWS cannot be made subjects of retribution – I do not contest that, as Sparrow posits, there is a necessary connection between retribution and suffering/deprivation, although some might be willing to do so[3]. Yet retribution is usually necessarily tangled with deterrence. All presently existing deterrence is based, after all, on some experience of suffering. Shame, guilt, self-contempt, regret and loneliness are as much feelings with a distinct, deeply unpleasant phenomenal quality to them as are the varieties of physical pain. Yet it is not necessarily so, even if it is so necessarily in humans. If one were to demonstrate that successful deterrence can be decoupled from retribution (a possibility that Sparrow does not seem to envision), then the machines incapacity for feeling would no longer stand in the way of the most important function of punishment being achieved. Thus the fact that they cannot be subjects of retribution, that is, inflicted predictable suffering upon as a response for criminal behavior, would lose much if not all of its significance.

For successful deterrence can also result from a threat to frustrate an agent’s goals and desires. The fact that such frustration is accompanied by unpleasant qualitative experiences in humans does not mean that these experiences are necessary to influence any agent’s behavior, or that the agent incapable of phenomenal experience cannot receive a negative incentive. Any artificial system useful in combat would work toward a certain purpose, and would avoid behaviors that make it less likely to achieve that purpose. As long as human commanders retain the ability to set and reset tasks for a LAWS – i.e., are genuinely in command of the system – and as long as these tasks can be sufficiently complex, that is, require fulfillment of several conditions for their completion, human commanders can either deter certain behaviors or prevent them outright.

Let us imagine two different types of such complex tasks, as exemplified by the following orders: 1) “clear this house of armed opposition without killing any unarmed humans; 2) “clear this house of armed opposition without killing more than three unarmed humans”. If the system is able to fully comprehend the meaning of order 1), then it is not only deterred, but effectively prevented from targeting unarmed humans *provided that* executing this order is its only purpose at the moment and that it understands that fulfilling only one of the two conditions is as much of a failure to execute the order as is fulfilling none of the two (a basic logical principle easy to drill into a machine yet sometimes hard to drill into a human mind). Comprehension and commitment to 1) would achieve perfect and assured compliance of the kind impossible in a human soldier. The fact that such compliance would be achieved through

[2] It may be impossible to reprogram some black-box algorithms and permanent decommission may be the only legally and morally acceptable choice, yet this is a flaw that might be acceptable to users.

[3] I thank an anonymous reviewer for bringing my attention to this possibility.

prevention and not through deterrence-by-threat-of-punishment would be of no ethical or legal importance.

Order 2), in contrast, would not prevent the occurrence of civilian casualties, yet it would powerfully disincline the machine away from causing them (and it would effectively prevent it from causing more than three fatalities). The system would still sometimes choose to cause these casualties, but it would not do so wantonly, since it would know that by doing so it limits its options for future action (since the fourth casualty simply mustn't be caused) and so decreases the likelihood that it will fulfill its purpose[4]. Thus it is able to cause civilian casualties but will, on many occasions, choose not to do so because of the negative incentive build into its orders. Thus 2) offers an example of deterrence *par excellence*. A modification of 2) making use of probabilities could be possible if the purpose set for the robot became more meta-oriented. Just like a dog may simply want to be called a good boy, rather than to become a successful ball-fetcher, a robot may be made to prefer its mission being declared a success by a human reviewer, rather than simply fulfill a number of pre-set conditions[5]. Imagine 2b) "clear the house of armed opposition. If any unarmed humans are killed, your mission will be declared a failure with a probability of 90%." The machine issued with 2b) will not kill an unarmed person unless it judges such an action necessary to prevent the likelihood of mission success from falling below 10%. This is a rather robust level of disinclination-short-of-prevention, best classified as deterrence of another kind than the one offered by a currently existing system of military justice, yet no less valid for that[6].

Of course the level of artificial intelligence required for comprehension and execution of complex orders like 1), 2) and 2b) is quite high and unlikely to be met by the first generation of autonomous machines. These systems are more likely to comprehend only much simpler orders, such as 3) "Shoot every human in the house", as they may not be able to distinguish morally relevant features of a situation, e.g., tell an armed person from an unarmed person. Such simple systems cannot be deterred any more than grenades can be. In fact they would differ from grenades only in their efficacy and sophistication, yet not morally. Completely oblivious to the ethical aspects of the situation, they would constitute just another category of highly indiscriminate weaponry to be used only in specific circumstances, in which their moral obliviousness would not be problematic. The commanders deciding to use them would bear the full blame for all the consequences, and they would not present military ethicist with any novel problems. Nor would their malfunctions do so (Simpson & Mueller 2015). The ethical and legal

[4] This may no longer be true close to the end of the systems mission. Imagine that a system is tasked with clearing a four-room house, each room containing three civilians. If a system following 2) is successful in clearing the first three rooms without any casualties, it may then go on and simply throw some grenades into the last room, which would be wanton and disproportional given its demonstrated ability to clear rooms without killing civilians. This shows human supervision, consisting in the ability to reset the systems orders with the progress of events, is always desirable as far as it can be achieved. What it does not show is that using the robot capable of such moral failings should be prohibited – it may still be the best among flawed alternatives.

[5] Of course such a solution raises a prospect that the machine could engage in wire-heading – trying to hack its own reward path – yet I do not believe it would be a serious danger at the level of AI sophistication envisioned in this article.

[6] The use of such machines would still be subject to scrutiny on the basis of the Principle of Proportionality. Yet it would be the humans drafting the machines' rules of engagement who would be to blame for disproportionate actions.

tools already developed to deal with indiscriminate weaponry, weapon testing and manufacturers' responsibility would be enough for that purpose (Lucas 2013).

I have shown that both deterrence and substitution of deterrence by more effective prevention is possible for LAWS without making them sentient or attempting to punish them. This leaves only the retributive aspect of punishment to account for. Yet it is doubtful whether the retributive aspect even needs to be accounted for, if it becomes possible to completely decouple it from the other three. If deterrence, incapacitation and rehabilitation were possible without infliction of any suffering, should we not enthusiastically embrace such a possibility? My intuition is that we should. More importantly, LOAC and Just War Theory are indeed focused on preventing unnecessary harm and limiting the tragedy that is war, rather than on delivering cosmic justice to everyone involved. The Principle of Moral Equality of Combatants, still the centerpiece of Laws of Armed Conflict, is a paradigmatic example of this rather pragmatic and consequentialist orientation (McMahan 2009, 108-110).

It is highly doubtful that shorn of its instrumental value as a contributor to deterrence, retributive punishment retains much axiological weight (Caruso 2018). An argument based solely on the value of retribution itself would not possess anywhere near the level of credence of Sparrow's original argument, rooted in a joint value of all four punishment functions. And even if one was to attach inherent value to retributive punishment, any entitlement rooted in this value would not carry a weight comparable to other rights and values at stake. LAWS' non-trivial potential for limiting the suffering involved in war by, among others, decreasing casualties among human soldiers (a possibility taken seriously by Sparrow himself) clearly outweighs any moral upside of retribution. Not every type of LAWS, and not every use of LAWS, will live up to this potential – perhaps most will not. Yet if even a single class of LAWS, used in a single specific way by a single specific actor will happen to reduce these harms, then their inability to be subjected to retribution will fail to ground moral opposition to the use of this specific type of LAWS[7].

I have demonstrated that restrictions placed on LAWS can equal and most probably supersede the benefits of three out of four aspects of punishment as administered to humans – incapacitation, rehabilitation, and deterrence. The absence of the fourth aspect – retribution - does not seem to carry much moral weight and may even be welcome on ethical grounds. Even if there was some inherent value in retribution, it would not be the kind of value that could justify eschewing life-saving advances in autonomous technologies. This undermines Sparrow's argument from problems with responsibility attribution. While other ethical problems with policing the use of LAWS remain to be dealt with and further research into the issue is urgently needed, these weapons are not intrinsically evil or fundamentally incapable

[7] It is worth noting that the dubious value of delivering retribution is quite routinely outweighed during post-war reconciliation. When confronted with the choice between granting immunity to perpetrators of war crimes as a price of lasting peace and continuation of hostilities or undermining peace arrangements nations routinely choose the former, the immunity granted to Japanese Emperor Hirohito by the Allies being the best historical example. Similar choices were made during various "velvet revolutions" to facilitate peaceful transfer of power by authoritarian regimes guilty of human rights violations. It is hard to view these choices. as morally unacceptable, even though in these cases other aspects of punishment are not as clearly untangled from retribution.

of compliance with LOAC, and certainly not on the grounds that their illegal actions would necessarily go unpunished.

## Bibliography

Caruso, Gregg D. "Justice without Retribution: An Epistemic Argument against Retributive Criminal Punishment." *Neuroethics* (2018): 1-16.

Demleitner, Nora V. 2014. “Types of Punishment.” In *Oxford Handbook of Criminal Law*, Edited by Markus D. Dubber and Tatjana Hörnle, 941-963. Oxford: Oxford University Press.

Lucas, George R. Jr. 2013. “Engineering, Ethics, and Industry: The Moral Challenges of Lethal Autonomy.” In *Killing by Remote Control*, Edited by Bradley J. Strawser, 211-228. New York: Oxford University Press.

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

Simpson, Thomas W. and Müeller, Vincent. 2015. “Just War and Robots’ Killings.” *The Philosophical Quarterly* 66 (263): 302-322.

Sparrow, Robert. 2007. “Killer Robots.” *Journal of Applied Philosophy* 24 (1): 62-77.