Ethics of artificial intelligence: virtue ethics as a solution to artificial moral reasoning in the context of lethal autonomous weapon systems

Karabo Maiyane[1][0000-0003-2953-2871]

[1] University of Pretoria, Hatfield, Pretoria 0002, South Africa
samuelmaiyane@gmail.com

Abstract. In 2013, a coalition of Non-governmental organisations launched a campaign called the Campaign to Stop Killer Robots [1]. The ongoing aim of the campaign is to raise awareness about the dangers of developing and deploying Autonomous Weapons Systems (AWS). They are calling for the ban of such weapon systems. AWS are any weapon systems that, once activated, can select (i.e. search for or detect, track, select) and engage (use force against, neutralise, damage or destroy) targets without further human intervention [2]. In this campaign, they raise a series of moral, legal and political concerns regarding the development and use of such systems. Amanda Sharkey [3] summarises these arguments into three categories: "arguments based on technology and the ability of AWS to conform to international humanitarian law; (ii) deontological arguments based on the need for human judgement and meaningful human control, including arguments based on human dignity; (iii) consequentialist reasons about their effects on global stability and the likelihood of going to war".

In this talk, I will be focusing on the arguments based on technology and the ability for AWS to conform to international humanitarian law. Specifically, I will be addressing the following questions: *What moral status should AWS hold? How do we assign responsibility and accountability in the event of transgressions? Can AWS programming comply with laws of war (IHL)? For AWS to be able to comply with IHL, how would they have to be programmed?*

On the question of the moral status of AWS, I argue that AWS, by virtue of their definition and espoused functions, can no longer be considered as mere weapons and thus *instruments* of war. As such, in the context of warfare, they should be considered as *combatants*, meaning that their moral status should be that of *moral agents*. Ascribing AWS a moral status places us at a strategic position with regards to responding to the critique of whether AWS will be able to conform with IHL. Knowing that AWS are now considered as combatants means that we are now aware of which parts of IHL they must specifically comply with. For example, as combatants they would have to comply with *jus in bello* principles such as: discrimination and non-combatant immunity and proportionality, [see [4], [5], [6], [7]]. It also means that with regards to responsibility, they can now be held responsible the same way other combatants are. Knowing which parts of IHL AWS must comply with, what follows is to resolve the questions on how AWS must be programmed to ensure such compliance.

Regarding programming morality for artificial moral agents (AMA's), Wallach and Allen [8] published a book titled: *Moral Machines: Teaching Robots Right from Wrong*. In this book they explore different programming possibilities for AMA's. They make a case for both top down and bottom up approaches, and argue that both such systems

have shortcomings: top down in that explicit procedural systems don't work in all situations, especially in warfare where some situations cannot be anticipated; and bottom up in that learning systems need the position from which to learn from. They argue that "If neither a pure top-down approach nor a bottom-up approach is fully adequate for the design of effective AMAs, then some hybrid will be necessary" [8, p. 117]. This is an approach that combines both top-down and bottom up approaches. Many [see [4], [5], [8], [9]] who argue in favour of hybrid approaches consider virtue ethics as an important normative approach. Specifically, the focus is on Aristotle's conception of virtue as twofold, moral and intellectual. For Aristotle [10], intellectual virtues can be *taught*, and moral virtues can be cultivated through *habit* over time. It is because of this distinction he makes that researchers in the field find his conception of virtue attractive for hybrid approaches. Wallach and Allen [8] argue that intellectual virtues can be programmed using a top-down programming, whilst moral virtues can come about as a result of a "bottom up process of learning and discovery by an individual through practice".

I argue that a hybrid program, using a virtue ethics normative approach, would be the solution for moral reasoning in the case of AWS. I wish to clarify here that by a virtue approach I do not mean programming specific virtues into AWS; I mean using the framework that Aristotle argues should be followed by human agents in terms of cultivating virtues. I argue this because for Aristotle, virtues are those characteristics that makes one good at what they do. Thus, their acquisition is through learning how to be good at that which one does, by doing it constantly – from being taught rules (intellectual virtues), to learning them by continuous practice over time (moral virtues). I will illustrate that Aristotle's conception of how virtues are acquired offers a good framework for building an architecture of an autonomous moral agent, especially one that operates under defined contexts such as AWS in warfare. In this sense the top down approach would be used to program IHL while the bottom up approach would use machine learning as learning program to train AWS.

I am not arguing that a virtues-based architecture is the only one that can work in programming morality in all AMAs, only that in the case of AWS in warfare it is ideal. This is because such an architecture allows for universal programming of rules (IHL) on the one hand but accepts dynamic applicability (applicability based on context and agent) on the other. Which is what is required of combatants in warfare.

Keywords: Autonomous Weapon Systems, Artificial Intelligence, Virtue ethics, Moral agency, Just war theory.

References

1. The Campaign to Stop Killer Robots, https://www.stopkillerrobots.org/learn/#problem, last accessed 2019/11/25.
2. International Committee of the Red Cross, https://www.icrc.org/en/publication/4283-autonomous-weapons-systems, last accessed 2019/11/25.
3. Sharkey, A.: Autonomous weapons systems, killer robots and human dignity. Ethics and Information Technology 21(2), 75-87 (2019).
4. Lin, P., Bekey, G., Abney, K.: Autonomous Military Robots: Risk, Ethics, and Desing. Cal Poly (2008).
5. Abney, K. Autonomous Robots and the Future of Just War Theory. In: Routledge Handbook of Ethics and War: Just War Theory in the 21st Century Fritz, A. (eds.) pp. 338 – 351 Taylor and Francis (2013).
6. Orend, B. The Morality of War. 2nd edn. Broadview Press:Toronto (2013).
7. Lazar, S. The Stanford Encyclopeadia of Philosophy, https://plato.stanford.edu/archives/spr2017/entries/war/, last accessed 2019/10/5.
8. Wallach, W., Allen, C.: Moral Machine: Teaching Robots Right From Wrong, Oxford University Press, New York (2009).
9. Casebeer, W.D.: Natural Ethical Facts: Evolution, Connectionism, and Moral Cognition, MIT press,Cambridge (2003).
10. Aristotle. Nicomachean Ethics Bartlett, R.C., Collins, S.D. (Eds.), The University of Chicago Press, Chicago (2011).