©*International Journal of Applied Philosophy*. ISSN 0739-098X.
doi:
Online First:

# Employing Lethal Autonomous Weapon Systems: Ethical Issues in the Use of Artificial Intelligence in Military Leadership

**Matti Häyry**
***Aalto University School of Business***
***and National Defence University***

**ABSTRACT:** The ethics of warfare and military leadership must pay attention to the rapidly increasing use of artificial intelligence and machines. Who is responsible for the decisions made by a machine? Do machines make decisions? May they make them? These issues are of particular interest in the context of Lethal Autonomous Weapon Systems (LAWS). Are they autonomous or just automated? Do they violate the international humanitarian law which requires that humans must always be responsible for the use of lethal force and for the assessment that civilian casualties are proportionate to the military goals? The article analyses relevant documents, opinions, government positions, and commentaries using the methods of applied ethics. The main conceptual finding is that the definition of autonomy depends on what the one presenting it seeks to support. Those who want to use lethal autonomous weapons systems call them by another name, say, automated instead of autonomous. They impose standards on autonomy that machines do not meet, such as moral agency. Those who wish to ban the use of lethal autonomous weapon systems define them much less broadly and do not require them to do much more than to be a self-standing part of the causal chain.The article's argument is that the question of responsibility is most naturally perceived by abandoning the most controversial philosophical considerations and simply stating that an individual or a group of people is always responsible for the creation of the equipment they produce and use. This does not mean that those who press the button, or their immediate superiors, are to blame. They are doing their jobs in a system. The ones responsible can probably be found in higher military leadership, in political decision-makers who dictate their goals, and, at least in democracies, in the citizens who have chosen their political decision-makers.

## THE TASK

Military leadership already relies, to an extent, on machines and autonomous systems, some based on artificial intelligence (AI), that appear to make

decisions for human beings. In the future, the use of such devices can increase considerably. Some of the devices and systems employing them are involved in decisions that do not belong exclusively to the military, or to military leadership. Lethal Autonomous Weapon Systems (LAWS) provide a major instance of an almost exclusively military application (although at least law enforcement could use them, too). The task here is to examine what ethical issues we should account for in the use of AI-assisted LAWS and other relevantly similar systems.

## THE GENERAL BACKGROUND ETHICS AND POPULAR EXTENSIONS

Isaac Asimov's three "laws of robotics" provide, despite their fictional origin in the short story *I, Robot* (1950), a good starting point for defining the ethics of robots and AI. The intuitively appealing laws presented by Asimov were:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The British Engineering and Physical Sciences Research Council (EPSRC) and Arts and Humanities Research Council (AHRC) principles of robotics[1] and Satya Nadella's laws[2] give accounts that are more detailed. They include elements such as the prohibition of killing; a requirement of traceable human responsibility; an emphasis on safety and security; a call against making robots too humanlike or otherwise too likeable to humans; a demand for transparency and understandability; and respect for human autonomy, dignity, privacy, and equality. The application of these general rules is not, of course, limited to the military use of AI or AI-assisted machines. These, too, fall, however, under their domain.

## HOW DO LAWS FARE WITH GENERAL ETHICS?

Whichever set of principles we choose, LAWS appear to stumble on the first hurdle. They foreseeably, in their normal and intended use, injure or kill human beings. This is probably the main reason for attempts to ban them in the United Nations (UN),[3] in the European Parliament (EP), and among AI developers. It also lies behind rejections by concerned academics and the Roman Catholic Church. (More on this shortly.)

Since warfare in general, however, involves injuring and killing human beings, traditional just-war theories can salvage LAWS, albeit with major caveats. As long as an activity is primarily intended for a good purpose (e.g., the defence of a nation's citizens against aggressors), it is, according to the Doctrine of Double Effect and its allies, sometimes justified even if it also has a bad effect.[4] [5] The conditions for this include that the bad effect (injury and death especially among civilians) is not actively intended (if citizens could be protected by other means, injury and death would not be caused) and that the bad effect is proportionate to the good.

When we apply these *jus ad bellum* considerations to the *jus in bellum* legitimation of LAWS, two specifications emerge:

- LAWS must be able to distinguish, as reliably as possible, combatants from non-combatants.
- The damage caused by LAWS to non-combatants must be proportional to the military achievement.

These are also required by International Humanitarian Law (IHL). The Ottawa Treaty's ban of anti-personnel landmines is based on such considerations, although mines and mine fields not involving the use of AI are not as such classified as LAWS.

## FROM AN EXTENDED CONCERN TO SPECIFIC QUESTIONS

The Roman Catholic Church opposes the use of LAWS due to their feared impact on the humanity of warfare. In a presentation of the Church views, Alice de la Rochefoucauld introduced the unethical nature of LAWS by a series of questions:

> Are machines capable of replacing the human person in decisions over life and death and is this compatible with International Humanitarian Law? Can machines be responsible for the violations of international law? Ethically, can a machine replace the human capacity of moral reasoning?[6]

The unequivocal answer to all these is "No," and the corollary is that accepting machine involvement in the form of LAWS would depersonalise and dehumanise warfare. Stripping this of its metaphysical baggage (the romanticised "humanity" of traditional warfare), the concrete concern remaining is that even more innocent civilians will suffer, and nobody can be held responsible.

Put like this, the cure is obvious, at least theoretically. The humans responsible for producing and deploying LAWS are responsible for the machine-made decisions over life and death and the possible violations of international law. It does not matter where the human is relative to the loop—in, on, or out of it. Machines are not capable of moral reasoning. Humans are. The last human making the meaningful, autonomous decision is responsible. Or can it be that simple? Do we need to take a closer look at de la Rochefoucauld's questions and treat them as literal instead of rhetoric?

## HUMAN AUTONOMY AND AUTHORITY

Moral philosophy recognises two main forms of *human* autonomy. According to an individualistic account, decisions are autonomous if and only if agents make them based on their own deliberations, convictions, and values, unencumbered by external social or cultural forces. According to more relational accounts, decisions are autonomous if and only if agents make them based on their own deliberations and convictions, observing the right social and communal norms and values.[7] Whether or not autonomous decisions must also be well-informed is debatable. At least valid consent procedures in healthcare and scientific research separate the requirements, and individuals make valid decisions only when they make

them freely (they are not forced or coerced) *and* informedly and autonomously. This suggests that autonomy does not always imply extensive knowledge.

Military leadership decisions typically involve two kinds of authority.[8] The first is *expertise* authority, based on knowledge. On bigger-unit levels, the officers in charge of artillery, engineering, signals and communication, and so on provide the exact information and knowledge needed for an operative decision. The Commanding Officer (CO) then combines (with the help of the staff) the information from different sources and makes (autonomously) a choice that becomes binding by the CO's organisational *position* authority. The picture is slightly different on smaller-unit levels, and changes radically with the Strategic Corporal making split-second choices that may have global consequences.[9] [10] The expertise aspect is still there, but in a considerably smaller role than in the higher-level cabinet resolutions.

## MACHINE AUTONOMY—INCLUSIVE

The autonomy of weapons and weapon systems has different interpretations, which are dictated by the normative views and practical needs of the ones formulating the definitions. Those who want bans and restrictions go for inclusiveness (assign autonomy to as many weapon systems as possible), while those who oppose strict constraints rely on exclusiveness (limit autonomy to fewer systems).

Mark Gubrud, a peace researcher, provides an example of inclusiveness.[11] As a descriptive starting point, he uses the United States Department of Defense definition, according to which an autonomous weapon system (AWS) is:

> A weapon system that, once activated, can select and engage targets without further intervention by a human operator.[12]

Weapon system autonomy here means that after the human decision has been made, the machine "takes over" and makes new "selecting" and "engaging" decisions over which human operators have no control. We could say, as I suggested above, that the last human in the chain is responsible. Gubrud does not take this view. Instead, he seems to think that human responsibility evaporates when the machine makes choices on its own. Since he also believes that "Weapons and conflict must always be under human control," his normative conclusion is clear. There should be a general ban on AWS, and anyone who intends to develop and use weapon systems resembling them must find separate justifications for their choice.

Gubrud goes on to argue that the attempts by the Department of Defense to distinguish between (acceptable) Semi-Autonomous Weapon Systems (SAWS) and (unacceptable) Fully Autonomous Weapon Systems (FAWS or AWS) fails. In some SAWS, humans only have to make the final "engage" decision. Gubrud notes that such SAWS could be converted into FAWS far too easily for the distinction to be safe. In "fire and forget" systems, even that minimal human involvement has been removed, as after launching the machine selects and engages the target by itself. The Department of Defense's view is that the calculations leading to the launch constitute the critical human decision, Gubrud disagrees, and from this

point on the controversy becomes semantic. Are the SAWS "select" and "engage" functions autonomous machine decisions or not? One possibility is that they are, in a sense, autonomous (machine working without external controls) but not really decisions (in the sense that incurs responsibility).

## MACHINE AUTONOMY—EXCLUSIVE

The United Kingdom Ministry of Defence sets the standard higher in its definition of weapons system autonomy:

> An autonomous system is capable of understanding higher-level intent and direction. From this understanding and its perception of its environment, such a system is able to take appropriate action to bring about a desired state. It is capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control, although these may still be present. Although the overall activity of an autonomous unmanned aircraft will be predictable, individual actions may not be.[13]

The last sentence about predictability contrasts autonomous systems with automated ones. The latter produce only outcomes that we know if we know what rules the system is programmed to follow. Autonomous systems would go beyond this and produce unpredictable outcomes, which would then have to be attributed to machine intelligence. If we hold on to the tenet that machines should not make choices that involve moral and legal responsibility, genuine AWS would have to be banned. By defining AWS-looking systems as remote automated ones, however, the ban can be averted. This is what the Ministry of Defence seems to be doing by its classification of unmanned aircraft systems as "remote automated," recognising that this is their interpretation of international law yet holding on to the reading as legitimate.

The opening sentence of the definition supports the idea by requiring that truly autonomous systems are "capable of understanding higher-level intent and direction." Much depends on what "understanding" ja "higher-level intent and direction" are supposed to mean, but they sound ominously demanding, requiring qualities approaching human autonomy. This sets the threshold so high that no machine without a soul or a mind or some equivalent can reach it.

## SELECT, ENGAGE, AND MACHINE AUTHORITY

Both critics and advocates seem to credit LAWS, at least in theory, with both kinds of authority that are involved in the CO's decision. After the machines have left human control, they select their targets independently, thereby exercising *expertise authority*. Once the system has concluded its selection, it then appears to make a further decision to engage with the target. This second step, the machine exercising *position authority*, is the ethically alarming one, but probably also an anthropomorphic misinterpretation that needs to be debugged.

Expertise authority in the first step, "select," is something that we should gladly delegate to well-functioning machines. Which one would we prefer—an explosive artillery projectile sent responsibly by a human operator or a choice-

making out-of-human-control unmanned aircraft with equivalent firepower? When the artillery shot reaches me, a non-combatant, it explodes and kills or maims me. The drone, properly programmed, will independently of its human masters make one last check, identify my status correctly, and go away. The letter (if not the spirit) of IHL could be better served by good machines than by humans.

Position authority in the second step, "engage," is what ethicists seem to be wary about. Can machines replace human persons in decisions over life and death, as de la Rochefoucauld puts the matter. She does not, apparently, mean the kind of assessment that the drone in my example makes, but something more. What more is there, though? Human emotion, perhaps. A person might pity the people about to be harmed and decide to abort the mission. Then again, a person might be outraged about something and engage. Whatever the case, ethicists appear to want a human act of will incurring moral responsibility to be present, and that makes the matter metaphysically muddled.

## UTILITARIAN ELECTRIC FENCE ETHICS AS AN ALTERNATIVE

The need for the moral-responsibility-incurring act of will arises from the Aristotelian, Kantian, natural law, and human rights ethics that form the ideological background of IHL.[14] An alternative exists, a utilitarian or pragmatic alternative that focuses on consequences instead of motives and intentions.[15] If we take IHL to aim at reducing non-combatant damage caused by LAWS (among other warfare technologies and practices), responsibility could be defined in terms of deterrence. Legal rights and duties (according to this view) are fictions designed for a purpose, so the question is, how should we assign responsibility to optimise the deterrence against using LAWS on non-combatants? Punish the last humans in the chain? Possible, but probably unfair, because their positions in the organisation more than likely coerce them to push the button. Punish their superior officers? Possible, but the same consideration partly applies. Punish the high command? Possible, but the same consideration may apply. Punish political leaders? Possible, but who is in a position to punish them? International law is not perfectly enforced.

Outside the chain of command—from going to war to launching the lethal machine—we find yet other alternatives. The engineers who devised the algorithm? The voters who put the political leaders into their position? The foreign aggressors who necessitated the war? The global capitalists whose actions and inactions created the circumstances in which war was inevitable? The consumers who by their choices promoted global capitalism? The possibilities are myriad, but if we keep the aim in mind, the identification of the optimal deterrence should be reachable by empirical investigation. Note that this (utilitarian) theory of punishment is not necessarily concerned about guilty minds, deeper moral responsibilities, or even innocence. Legal sanctions simply should be defined so that they are enforceable and effective and deter non-combatant damage. Perhaps that would encourage the development of AWS that select discriminately. We could then see engagement simply as the last stage of selection as far as the machine is concerned and ignore philosophical conundrums about that phase.

## NON-COMBATANTS, COMBATANTS, PROPORTIONALITY, AND THE NECESSITY OF KILLING

Some outside-the-box questions remain. Why concentrate exclusively on the protection of non-combatants? They are vulnerable and, in many cases, relatively innocent, but so are many combatants. In addition, non-combatants can be a part of the war effort, sometimes willingly. Why should they, then, be exempt from the damage (apart from the IHL say-so)?[16] [17]

This re-raises the question of proportionality, the second IHL requirement. When is non-combatant damage proportional to the military achievement? When is combatant damage proportional to it? What is military achievement? When and how is it commensurable with the death and injury inflicted?

Again, this is a concern that could be addressed by down-to-earth pragmatic thinking. One line of thought could then be to challenge the killing aspect. Why is the arms industry, with the support of governments, developing LAWS when they could be developing INLAWS, PINLAWS, or TINLAWS? By these abbreviations I mean Incapacitating Non-Lethal Autonomous Weapon Systems, Permanently Incapacitating Non-Lethal Autonomous Weapon Systems, and Temporarily Incapacitating Non-Lethal Autonomous Weapon Systems. The last category sounds particularly appealing, despite the challenges posed to it by existing regulations against biological and chemical warfare.[18] If the aim of international legislation is to minimise death and permanent damage to non-combatants and combatants alike, and if warfare is still deemed to be necessary, these regulations should perhaps be reconsidered.

## SUMMARY AND CONCLUSIONS

To summarise, some ethicists have been worried that AI-assisted LAWS and other AWS make decisions that only human beings ought to make. By using such technology not only are we playing God[19] ourselves but also enabling machines to do the same. The view emerging in the analysis above is, however, that this may be an exaggerated metaphysical concern. Machines make selection decisions and in some sense exercise expertise authority independently of humans, but this is not necessarily alarming. The engage "decisions" that machines make are probably best seen as final selection decisions, not as expressions of position authority, and they do not contain mysterious displays of intention or acts of will. These can only be attributed to human choices, and they are what make humans morally responsible for their actions and inactions. This means that the machine is never responsible for the damage it causes, human beings somewhere in the chain are.

With this philosophical problem solved, we are still left with the legal issue. IHL requires LAWS to distinguish between combatants and non-combatants and engage only if the damage to non-combatants is proportional to the military achievement. If we assume the utilitarian model of punishment as deterrence, whom should we punish and for what crime? The spirit of IHL is clearly that two types of crime are possible, both in the decisions to develop specific kinds of LAWS. First, LAWS must identify non-combatants, perhaps especially innocent ones.* Secondly, LAWS must not engage without a completed and accepted

proportionality assessment. To develop and employ LAWS that fail to do one of these is a crime.

Not all countries recognise these crimes, because they use their own definitions and choose to interpret IHL in a way that supports their own use of autonomous or automated weapon systems. Not much can be done about this, as there is, in the absence of UN decisions, no international power that could correct them. If an agreement is eventually forged, IHL could dictate that the political, policy, and business decision makers who order, design, manufacture, and use insufficiently discriminating LAWS should be held liable.

## NOTES

* Interestingly, an evaluation of innocence, far-fetched as it may sound, could now or soon be completed by using information about the potential target's net presence and activities. We can be categorised for commercial purposes already, so it should not be difficult to categorise us for military/humanitarian ones.

## ACKNOWLEDGEMENTS

This article was originally commissioned by the National Defence University and the Finnish Defence Forces Defence Command; and written as a background paper for the Multinational Capability Development Campaign Future Leadership. The Academy of Finland (project Bioeconomy and Justice, SA 307467) and the Finnish Ministry of Agriculture and Forestry (projects The Role of Justice in Decision Making Concerning Bioeconomy and A Just Management Model for Systemic and Sustainable Shift Towards Bioeconomy) have supported my work. My thanks are due to all my sponsors.

## ENDNOTES

1. EPSRC—Engineering and Physical Sciences Research Council (2011). *Principles of Robotics*. https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

2. Satya Nadella, "The Partnership of the Future," *Slate* 28 (June 2016). https://slate.com/technology/2016/06/microsoft-ceo-satya-nadella-humans-and-a-i-can-work-together-to-solve-societys-challenges.html.

3. Damien Gayle, "UK, US and Russia among those opposing killer robot ban," *The Guardian* (29 March 2019). https://www.theguardian.com/science/2019/mar/29/uk-us-russia-opposing-killer-robot-ban-un-ai.

4. Alison McIntyre, "Doctrine of Double Effect," *The Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, 2019, https://plato.stanford.edu/archives/spr2019/entries/double-effect/.

5. Seth Lazar, "War," *The Stanford Encyclopedia of Philosophy*, ed. Edward Zalta, 2020. https://plato.stanford.edu/archives/spr2020/entries/war/.

6. Alice de la Rochefoucauld, "Introduction to the documents of the Holy See," *The Humanization of Robots and the Robotization of the Human Person: Ethical Reflections on Lethal Autonomous Weapons Systems and Augmented Soldiers*, ed. Alice de la Rochefoucauld and Stefano Saldi (Caritas in Veritate Foundation Working Papers 2017). http://www.fciv.org/downloads/WP9-Book.pdf.

7. Matti Häyry and Tuija Takala "Coercion," *Encyclopedia of Global Bioethics*, ed. Henk ten Have (Cham: Springer, 2016), 595–605.

8. Richard Flathman, *The Practice of Political Authority: Authority and the Authoritative* (Chicago, IL: University of Chicago Press, 1980).

9. C. C. Krulak CC. "The strategic corporal: Leadership in the three-block war," *Marine Corps Gazette* 83, no. 1 (1999): 23.

10. David Lovell and Deane-Peter Baker, *The Strategic Corporal Revisited: Challenges Facing Combatants in 21st Century Warfare* (Cape Town: UCT Press, 2017).

11. Mark Gubrud, Mark, "Autonomy without Mystery: Where *Do* You Draw the Line?" *1.0 Human* 9, May 2014. http://gubrud.net/?p=272.

12. DoD—Department of Defense, *Autonomy in Weapon Systems*, Directive Number 3000.09, November 21, 2012. Incorporating Change 1, May 8, 2017. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf.

13. MoD—Ministry of Defence, *Unmanned Aircraft Systems*, Joint Doctrine Publication 0–30.2 (JDP 0–30.2), dated August 2017.

14. Matti Häyry, "Doctrines and Dimensions of Justice: Their Historical Backgrounds and Ideological Underpinnings," *Cambridge Quarterly of Healthcare Ethics* 27, no. 2 (2018): 188–216, at 203–204.

15. Matti Häyry, *Liberal Utilitarianism and Applied Ethics* (London: Routledge, 1994), at 164–166.

16. Seth Lazar—see note 5.

17. Jeff McMahan, "Who Is Morally Liable to Be Killed in War?" *Analysis* 71, no. 3 (2011): 544–559.

18. Fritz Allhoff, "The Paradox of Nonlethal Weapons," *Stanford Law School Biosciences Blog* 10 March 2016. https://law.stanford.edu/2016/03/10/the-paradox-of-nonlethal-weapons/.

19. Matti Häyry, "Categorical objections to genetic engineering – A critique." *Ethics and Biotechnology*, ed. Anthony Dyson and John Harris (London: Routledge, 1994), 202–215.