

GC REAIM

GLOBAL COMMISSION ON RESPONSIBLE
AI IN THE MILITARY DOMAIN



Foreign, Commonwealth
& Development Office

GC REAIM Expert Policy Note Series

Collective Moral Responsibility, Institutionalisation and Lethal Autonomous Weapon Systems (LAWS)

Seumas Miller

April 2025

POWERED BY



The Hague Centre
for Strategic Studies

GC REAIM Expert Policy Note Series

Collective Moral Responsibility, Institutionalisation and Lethal Autonomous Weapon Systems (LAWS)

Author: Seumas Miller

April 2025

Cover photo: [Unsplash](#)

The Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) is an initiative of the Government of the Netherlands that was launched during the 2023 REAIM Summit on Responsible Artificial Intelligence in the Military Domain in The Hague. Upon request of the Dutch Ministry of Foreign Affairs, the Hague Centre for Strategic Studies acts as the Secretariat of the Commission.

The GC REAIM Expert Policy Note Series was funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom. GC REAIM Experts maintained full discretion over the topics covered by the Policy Notes. The contents of the GC REAIM Expert Policy Note series do not represent the views of the Global Commission as a whole. The Policy Notes are intended to highlight key issues related to the governance of AI in the military domain and provide policy recommendations.

© The Hague Centre for Strategic Studies. All rights reserved. No part of this report may be reproduced and/ or published in any form by print, photo print, microfilm or any other means without prior written permission from HCSS. All images are subject to the licenses of their respective owners

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies
Email: info@hcss.nl
Website: www.hcss.nl



**The Hague Centre
for Strategic Studies**

Introduction

There are multiple definitions of lethal autonomous weapons systems (LAWS). In this policy note they are simply (and somewhat loosely) defined as weapons that once programmed and activated can locate and identify, then track and destroy human and other targets without the further intervention of a human operator. However, note that some such systems may also be able to ‘learn’ in a manner that enables them to adjust their own internal states, and thereby adapt their functioning in response to changing circumstances in the environment in which they are deployed.¹ The responses of such ‘self-learning’ systems are not entirely preprogrammed and are therefore, other things being equal, not as predictable as systems without a self-learning capability. In the case of autonomous *lethal* weaponry, i.e., LAWS, this degree of unpredictability might be thought to exclude the possibility of meaningful human control.² Of course, the ‘doings’ of non-autonomous or, at least, semi-autonomous AI based weapons systems might also be unpredictable. However, the problem of unpredictability is likely to be more acute in the case of autonomous lethal weaponry.³

Note that this unpredictability does not exclude the possibility of ascribing moral responsibility; there is not a *moral* responsibility gap,⁴ even if there is a legal responsibility gap. For if the use of LAWS will have unpredictable outcomes including, potentially, massive loss of human life, then the authorities (e.g., politicians, commanders) who knew that this was the case, or should have known it, but who, nevertheless, made the decision to deploy these AWSs, are morally responsible if and when their decision results in unjust harmful, or otherwise bad, outcomes. Moreover, others who had this information, such as designers, manufacturers and operators, would also have a share in the collective (i.e., joint – see below) moral responsibility for the bad outcomes in question.⁵

Does the unpredictability attendant upon self-learning LAWS necessarily exclude the possibility of *meaningful* human control? Certainly, unpredictability can reduce the level of effective control. However, control is a matter of degree and the unpredictability in question does not necessarily extinguish human control. Accordingly, the question is whether the reduction in the level of control is morally unacceptable in the light of other relevant factors. So perhaps this argument for prohibiting LAWS should be recast as follows. There is a degree of human control but it is not meaningful because it is not morally acceptable. This argument is disputable, depending on, for instance, the likely

¹ Mariarosaria Taddeo, *The Ethics of Artificial Intelligence in Defence* (Oxford University Press, 2024), p. 173-176.

² Mariarosaria Taddeo, *The Ethics of Artificial Intelligence in Defence* (Oxford University Press, 2024), p. 198-203.

³ Nina Narodytska et al., ‘Verifying Properties of Binarized Deep Neural Networks’, *Proceedings of the AAAI Conference on Artificial Intelligence* 32, no. 1 (26 April 2018), <https://doi.org/10.1609/aaai.v32i1.12206>.

⁴ Robert Sparrow, ‘Killer Robots’, *Journal of Applied Philosophy* 24, no. 1 (2007): 62-77, <https://doi.org/10.1111/j.1468-5930.2007.00346.x>; Uwe Steinhoff, ‘Killing Them Safely: Extreme Asymmetry and Its Discontents’, in *Killing by Remote Control: The Ethics of an Unmanned Military*, ed. Jeff McMahan and Bradley Jay Strawser (Oxford University Press, 2013), 0, <https://doi.org/10.1093/acprof:oso/9780199926121.003.0009>.

⁵ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 279.

degree of unpredictability (perhaps low if, for instance, the LAWS is prevented from 'learning' and updating its functioning while undertaking a mission), the quantum of lethal force and consequent harm the weapon is capable of inflicting (e.g., the drone 'payload' might be quite small or the sentry robot might only be capable of the equivalent of small arms fire), the geographical reach, the quality of contextually relevant 'sensory' data, the effective firing range of the weapon (e.g., a drone travelling a few kilometres under favourable environmental conditions and delivering its payload at very short range), the proximity of civilians (e.g., it is 'trench warfare' and civilians have long since left the area) and, more problematically, what is morally at stake. In relation to what is morally at stake, we need to consider not only the military and, ultimately, political end in play, but also what the likely unintended consequences might be. Thus, a LAWS with a relatively small payload that is to be used against enemy combatants in the context of a genuine existential national threat such as, arguably, is currently confronting Ukraine as a result of Russia's invasion of its territory, might well be morally justified. By contrast, deploying a nuclear armed LAWS the use of which might result in a nuclear war that destroys much of humankind would be unconscionable.

The main focus in this paper is with the deployment and especially use (as opposed to, for instance, the design) of LAWS (which is not to say that questions of deployment and use can be decided independently of design, legal and other institutional arrangements, etc.).⁶ The two substantive points to be kept in mind in respect of use are that, firstly, target selection (in the context of some strategy or tactic, e.g. targeted killing of members of certain categories of enemy combatants, such as middle-ranking officers and above) and, secondly, engagement can be determined, in effect, by AI processes constitutive of an autonomous weapon.

It is by now generally agreed that LAWSs ought to be subject to control by their morally responsible human operators.⁷ This is in large part because only human operators can reasonably be expected to understand and comply with moral principles, including the laws of war (notwithstanding attempts to prove otherwise).⁸ Human beings are capable of being morally responsible for their actions, but AI enabled robots are not capable of being morally responsible for their 'doings'.

⁶ Filippo Santoni De Sio and Jeroen van den Hoven, 'Meaningful Human Control Over Autonomous Systems: A Philosophical Account', *Frontiers In Robotics and AI* 5, no. 15 (2018), <https://doi.org/10.3389/frobt.2018.00015>; Sten Allik et al., 'A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications', *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*, October 2024, 1–63.

⁷ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 271–283.

⁸ Ronald C. Arkin, 'The Case for Ethical Autonomy in Unmanned Systems', *Journal of Military Ethics* 9, no. 4 (December 2010): 332–41, <https://doi.org/10.1080/15027570.2010.536402>; Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 271–283.

Here the notion of meaningful human control (MHC) is relevant.⁹ However, the term “meaningful” is potentially ambiguous. For instance, it could be taken in the narrow sense of *effective* human control, or it could be taken in the wider sense of human control, the exercise of which consists in *morally correct* actions (or, at least, in the sense that the controller is a morally responsible human being). Both senses are no doubt legitimate (as others might be) but it is important to be clear which sense is being used in any given context.

A key question that now arises is that of moral responsibility; the moral responsibility of human beings with respect to the control of LAWS. There are three categories of responsibility directly relevant to our concerns in this paper.¹⁰ The first category is responsibility in the sense of *causal* responsibility. The bearers of causal responsibility are causal agents, i.e., entities that have causal powers and are responsive to causal factors. These include volcanoes and AI enabled robots, such as LAWS - as well as human beings. The causal agents in question here, such as human beings or AI enabled robots, are able to control other entities; they are controlling agents.¹¹

Note that machine control is a narrower notion than that of *human* control, given that human control implies the controller’s *understanding* of the nature and limits of its control over the controlled (including the moral significance of this control), the controller’s exercise of *free will* (according to some analysis of free will) and the controller’s responsiveness in the exercise of its control to reasons (qua reasons), including *moral reasons* (qua moral reasons). Note also that laws and regulations (e.g., International Humanitarian Law (IHL), target selection, mission to be undertaken, including the basic means as well as the mission end or goal, rules of engagement (ROE), standard operating procedures (SOPs) for weaponry and so on) are (at least to a considerable degree) applicable under human control (or, at least, ought to be). However, this human control is, typically, *joint* human control. It is control exercised by multiple human beings acting cooperatively (even if in the context of hierarchical government, military etc. institutional structures). Human beings jointly act to determine laws and regulations, military missions, ROEs, operating procedures for weaponry etc.

The second relevant category of responsibility is *moral* responsibility, i.e., responsibility for actions and outcomes that have moral significance. The bearers of moral responsibility are human beings, but not, for instance, volcanoes or AI enabled robots.

The third category is *institutional* responsibility (including, but not restricted to, legal responsibility). This is essentially responsibility possessed by a person in virtue of their institutional role.

⁹ Heather Roff and Richard Moyes, ‘Meaningful Human Control, Artificial Intelligence and Autonomous Weapons | Briefing Paper for Delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)’, 2016; Filippo Santoni De Sio and Jeroen van den Hoven, ‘Meaningful Human Control Over Autonomous Systems: A Philosophical Account’, *Frontiers In Robotics and AI* 5, no. 15 (2018), <https://doi.org/10.3389/frobt.2018.00015>.

¹⁰ John Martin Fischer, ed., *Moral Responsibility* (Ithaca, N.Y.: Cornell University Press, 1986); Ellen Frankel Paul, Fred Dycus Miller, and Jeffrey Paul, eds., *Responsibility* (Cambridge University Press, 1999).

¹¹ Stuart Jonathan Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, 1995); Sven Nyholm, ‘A New Control Problem? Humanoid Robots, Artificial Intelligence, and the Value of Control’, *AI and Ethics* 3 (2023): 1229–39, <https://doi.org/10.1007/s43681-022-00231-y>.

The institutional rights and duties constitutive of an institutional role, and conferred on the occupant of the role by virtue of his or her occupancy of the role, might *also* be moral rights and duties, e.g., the moral and institutional right of a combatant deliberately to use lethal force against enemy combatants (and not necessarily in self-defence, e.g., in an ambush). Importantly, many institutional rights and duties are grounded in moral rights and duties; the institutional rights and duties enshrine, concretise and are given direction by, the prior moral rights and duties. Therefore, as argued elsewhere,¹² it would be practically impossible for beings that are not adequately morally sentient to occupy the institutional roles in question. Accordingly, for this reason alone, it would not be possible or, at least, it would be extremely dangerous, to try to get AI enabled robots to fill institutional roles that involve moral decision-making and do so by conferring institutional rights and duties (including legal rights and duties) upon them.¹³ Thus, LAWSs do not have a moral right, let alone a moral duty, to kill enemy combatants and, therefore, ought not to occupy the institutional role of a combatant (or be treated as if they did occupy that role).¹⁴ Relatedly, AI enabled robots cannot be held organisationally liable or accountable for the bad outcomes of their 'doings'; they cannot be held morally responsible and they cannot be punished. Nor for the same reasons can they be held criminally liable. In addition, they cannot be held criminally liable since they do not have *mental* states, notably intentions (specifically, *mens rea*) but rather only *functional* states (analogous to the functional states of complex biological systems, such as the human immune system, that are responsive to detected inputs in carrying out their functions). Intentions are conceptually connected to other mental states, such as beliefs and desires, and are such that the agents possessed of them are, or can become, conscious of them; but AI enabled robots do not have these other mental states and lack consciousness. Moreover, unlike institutional entities granted legal personhood, such as corporations, AI enabled robots are not in part constituted by human persons whose actions are, in turn, constitutive of their (the robots') 'doings'.

There are two notions of *collective* (as opposed to merely individual) responsibility relevant to our concerns here, namely, collective moral responsibility and collective institutional responsibility. *Collective* moral responsibility mirrors individual moral responsibility. Collective moral responsibility is the moral responsibility that attaches to the members of structured and unstructured groups for their morally significant actions and omissions. Elsewhere, Miller has elaborated and defended a relational account of collective moral responsibility; specifically, that of collective responsibility as joint responsibility.¹⁵ In this view, collective responsibility is responsibility arising from joint actions and omissions.

Collective institutional responsibility is the institutional responsibility that attaches to members of a group of institutional actors who perform a joint action qua members of the institution in question. Consider, for instance, the following scenario involving a

¹² Seumas Miller, 'Robots, Institutional Roles and Joint Action: Some Key Ethical Issues', *Ethics and Information Technology* 27, no. 1 (21 December 2024): 10, <https://doi.org/10.1007/s10676-024-09816-z>.

¹³ Seumas Miller, 'Robots, Institutional Roles and Joint Action: Some Key Ethical Issues', *Ethics and Information Technology* 27, no. 1 (21 December 2024): 10, <https://doi.org/10.1007/s10676-024-09816-z>.

¹⁴ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 271-283.

¹⁵ Seumas Miller, 'Collective Moral Responsibility: An Individualist Account', *Midwest Studies In Philosophy* 30, no. 1 (1 September 2006): 176-93, <https://doi.org/10.1111/j.1475-4975.2006.00134.x>.

drone strike. Assume a soldier on the ground reports that a man is digging the ground at some distance from the soldier and in doing so laying an IED (improvised explosive device). A drone is dispatched to hover overhead, conduct surveillance and relay video imagery back to the drone crew base. This imagery consists of numerous 'close-ups' from various angles and is far more reliable than the initial sighting by the soldier. The imagery is analysed by members of the crew, and it is correctly judged that the man is, as suspected, a terrorist laying an IED. The commander of the drone crew gives the order to the operator of the drone to fire and he does so, killing the terrorist. The killing of the terrorist involves cooperation between the soldier on the ground and the members of the drone crew, including its commander, those who analyse the imagery and the operator of the drone; it is a joint action for which the participants have collective, i.e. joint, moral responsibility, and potentially also joint institutional responsibility (in addition to the individual institutional responsibility that each has). Importantly, (prospective) institutional responsibilities can be attached by design to institutional roles (and new institutional roles designed), including in relation to weaponry (e.g. in the form of standard operating procedures) in a manner that not only tracks prior moral responsibilities, including joint moral responsibilities, but to some extent creates new individual and joint moral responsibilities. Moreover, once created, military role occupants can be held retrospectively institutionally (individually and jointly) responsible and, therefore, institutionally accountable and liable for their failures and relevant disciplinary measures taken and, in some instances, civil costs or even criminal sanctions imposed.

Let us now turn directly to actual applications of LAWS.

LAWS Applications: Integrating Technological, Institutional and Moral Dimensions

Firstly, in relation to LAWS technology, we need to invoke the distinctions between ‘human in-the-loop’, ‘human on-the-loop’ and ‘human out-of-the-loop’ weaponry. It is only human out-of-the-loop weapons that are autonomous. We assume in what follows that lethal human out-of-the-loop weaponry is morally unacceptable for the reasons elaborated above inter alia (e.g., inconsistency with MHC).¹⁶ Moreover, we also assume that if a lethal human-in-the-loop weapon is fit for (morally acceptable) purpose in the combat conditions in question then it morally ought to be preferred to lethal human on-the-loop weapons. However, we further assume that there are some combat conditions in which only lethal human-on-the-loop weapons (but not lethal human in-the-loop weapons) are fit for purpose, militarily if not morally. In relation to both ‘self-learning’ lethal human in-the-loop and human on-the-loop weapons, we assume for the reasons given in section 1 that the problem of unpredictability is not so acute in and of itself as to remove the possibility of MHC of all of these weapons.

Secondly, in relation to institutionally-based military roles relevant to LAWS, we need to invoke the distinctions between analysts (target identification and selection), operators (firing weapons) and the commanders thereof; and the distinction between using lethal force against combatants in close proximity to civilians and not doing so. We also need to invoke distinctions between orderly jurisdictions in peacetime, theatres of war, and disorderly jurisdictions without effective law enforcement experiencing ongoing, serious, armed conflict (e.g., the FATA, the Federally Administered Autonomous Areas in Pakistan which have been the site of US drone strikes on terrorists).¹⁷ We assume that it would be morally unacceptable to use LAWSs in orderly jurisdictions in peacetime, i.e., for law enforcement purposes.¹⁸

Thirdly, we need to invoke a distinction between decision-making made prior to a combat mission and decision-making (or, in the case of LAWS, determinations of ‘doings’) once the mission is underway. For instance, the term ‘target selection’ might refer to a decision made prior to the commencement of a mission, e.g., the selection of Osama bin Laden as the target was made prior to embarking on the mission to capture or kill him; it is *mission prior* target selection. Let us refer to this as *prior* target selection. Another instance of prior target selection would be to program a LAWS to use lethal force against members of a set of persons, such as all enemy combatants in a given geographical area, G. We need to distinguish prior target selection from target selection during a mission

¹⁶ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 271-283.

¹⁷ Milton C. Regan, *Drone Strike: Analyzing the Impacts of Targeted Killing* (Cham, Switzerland: Palgrave Macmillan, 2022), <https://doi.org/10.1007/978-3-030-91119-5>.

¹⁸ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 78-157.

already underway, i.e. what can be referred to as *intra* mission target selection. Thus, once the mission to kill or capture Osama bin Laden was underway, then the identification of a person under observation as being in fact the target, Osama bin Laden, would be an example of *intra* mission target selection. Again, once the mission to kill enemy combatants in area A during time period T is underway, then the identification of a person as being a target because he is an enemy combatant in A during T would be an example of *intra* mission target selection. For the reasons elaborated, 'self-learning' LAWSs ought not necessarily to be prohibited.

Human in-the-Loop

The use of a human-in-the-loop, even if AI enabled, lethal weapon in a theatre of war in compliance with relevant IHL and the related *jus in bello* moral principles of Just War Theory that govern the use of force once warfighting is underway (e.g., the principles of proportionality and military necessity) is, at least in principle, morally permissible (or so it will be assumed here). Moreover, other things being equal and notwithstanding that the lethal weapon in question is AI enabled, the human controller, or rather controllers, e.g., commander and operator, are: (i) individually morally and individually (prospectively) responsible for performing their respective role-determined actions; (ii) jointly (prospectively) morally responsible for attacking, indeed, killing enemy combatants, and, potentially, also jointly (prospectively) institutionally responsible for this (even if the operator has diminished moral responsibility and, under some command and control institutional arrangements, no institutional responsibility). Human-in-the-loop use of LAWSs involves, firstly, a lethal autonomous weapon system, i.e., one that does not require human intervention in its prior or (more likely) intra target selection and/or does not require human intervention in the delivery of lethal force once the intra target selection has been made. However, in the case of a so-called human in-the loop LAWS, the actual use of the LAWS does in fact involve human intervention in both prior and intra target selection and in the delivery of lethal force once the target selection has been made. In the case of intra target selection, the LAWS selects a target for consideration by the human analyst. Here the location and identification of a target for consideration, e.g., enemy combatants in a well camouflaged vehicle somewhere in a very large, heavily wooded, geographical area is much more efficient and effective, let us assume, than the counterpart human process. However, the human analyst needs to verify that the selected target is in fact, an enemy combatant by, for instance, analysing additional, now close-up, visual imagery of the behaviour of the (suspected) enemy combatants etc. In relation to the delivery of lethal force, on the basis of the analyst's judgment, the commander instructs the human operator to initiate the process of delivering lethal force at the target by locking the AI-controlled weapon onto the target. Once locked onto its target, the machine controlled weapon is far superior to the same weapon controlled by a human operator, especially given that the target vehicle may be equipped with an evasive capability or a defensive weapon. Such a human in-the-loop LAWS might be notionally characterised as follows¹⁹ (although there are a variety of possible alternative institutional arrangements to the ones used here):

(A) Human-in-the-loop.

1. Intra-mission target selection (e.g., at the level of a battle or small unit engagement) is jointly undertaken by human controllers (analyst and commander) in accordance with their respective institutional roles and (i) in light of the tactics of their human commanders, relevant laws, ROE, SOPs etc., and (ii) on the basis of the target location, identification and selection determinations of the AI enabled data collection and analysis processes of LAWS, e.g., facial recognition software;

¹⁹ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 277-278.

2. In accordance with their respective institutional roles, human controllers (operator and commander) jointly initiate LAWS' lethal attacks and in doing so rely on the intra-mission target selection provided by the analysts.

The joint (prospective) institutional and joint (prospective) moral responsibilities pertaining to the use of the human in-the-loop lethal weapon are discharged, as are the constitutive individual (prospective) institutional and moral responsibilities by their respective role occupants, i.e. human analysts, operators and commanders. Moreover, in this institutional arrangement, the commanders provide the 'knowledge-to-kinetic action' (institutional) link between analysts and operators as well as overall institutional command and control. As such they have a degree of individual (prospective) institutional responsibility for ensuring that the intended just outcomes of the actions performed are realised (as well as for the performance of the actions themselves) and, given the moral significance of these outcomes, individual (prospective) moral responsibility (and, therefore, in due course individual retrospective moral responsibility). Moreover, under many institutional arrangements, commanders are very likely not only to have individual retrospective moral responsibility for outcomes (bad and good) but also individual retrospective *institutional* responsibility and associated institutional accountability and liability (for bad outcomes in particular).

The analysts involved in this morally significant activity are (human) controllers and are, therefore, also (prospectively) *morally* responsible and, in due course, retrospectively morally responsible and, potentially, institutionally retrospectively responsible, accountable and liable for their determinations. After all, their target selection is not determined by their commanders; rather the analysts are, or ought to be, (prospectively) institutionally responsible for providing objective, independent determinations (even if these are confirmed and acted upon by their commanders). Moreover, analysts ought to be able to be trained to resist any inclination to become over reliant on AI based intra target selections.

The operators involved are also (human) controllers and are, therefore, simultaneously (prospectively) *morally* responsible and, in due course, retrospectively morally responsible and, potentially, institutionally retrospectively responsible, accountable and liable for their kinetic actions (and the bad outcomes of these actions), notwithstanding that they are subordinates acting on the orders of their commander. After all, they freely choose to 'pull the trigger' and ought not to do so if, for instance, the action would be unlawful or there is insufficient time to defer to their commanders.

In light of the above discussion, we can conclude that analysts, operators and their commanders are jointly (prospectively) morally responsible (in various combinations) and, potentially, jointly (prospectively) institutionally responsible (in various combinations), for a number of different joint actions involved in the use of human in-the-loop lethal weapons. Moreover, they may well be jointly (prospectively and, in due course, retrospectively) morally responsible for the outcomes of these joint outcomes. If so, then there is an argument for institutional arrangements under which they are jointly (and not merely individually) prospectively and, in due course, retrospectively, institutionally responsible (and accountable and liable) for at least avoidable, very morally bad outcomes. Indeed, there might even be grounds for going further and ascribing regulatory or criminal liability to the institutions or institutional units per se (as

well as to individual human role occupants, as appropriate) on grounds of the collective (i.e., joint) moral responsibility of their individual human role occupants).²⁰

²⁰ Seumas Miller, 'Corporate Crime, The Excesses of the 80s and Collective Responsibility: An Ethical Perspective', *Australian Journal of Corporate Law* 5, no. 2 (1995): 39–51.

Human on-the-Loop

The lethal use of a human-on-the-loop AI enabled weapon (which is not also a lethal human in-the-loop weapon, at least for all practical purposes in the combat conditions in question) is also, we assume, in principle morally permissible, albeit only under certain conditions. Moreover, the human operator is, perhaps jointly with others (such as the analyst and the commander – see above), morally responsible, at least in principle, for the use of lethal force and its foreseeable consequences and, therefore, potentially institutionally responsible. However, these two propositions concerning human on-the-loop AI enabled lethal weaponry (LAWS) rely on a number of assumptions²¹ including an understanding of what counts as a human *on*-the-loop AI enabled weapon.

The lethal use of a human-on-the-loop AI enabled weapon (which is not also a lethal human in-the-loop weapon, at least for all practical purposes in the combat conditions in question) can be characterised as follows and, thus characterised, is also, we assume, in principle morally permissible. A human on-the-loop weapon is one in which the analyst (and perhaps the commander) has the ability to override the intra target selection and/or the operator (and perhaps the commander) has the ability to override the weapons system. The ability to override, in our favoured sense, is a species of *human meta control* such that the operator (in particular) has: (i) The ability to shut down the weapon; (ii) The ability to replace machine control of the weapon by human control of that weapon, e.g., if the weapon goes haywire; (iii) However, when under human control, the weapon's performance *qua weapon* (in the contexts of armed conflict for which it has been designed) is much inferior to its performance when under machine control, e.g., when under machine control it might be a much more efficient and effective destroyer of enemy combatants than if under human control but in contexts of armed conflict in which many civilians are present more likely to harm civilians than if under human control; (iv) Shutting down the weapon returns its human operator to the *ex ante* situation, i.e. the armed force using the weapon is no worse off than if their operator had not activated machine control.

In order to facilitate our somewhat theoretical discussion of human on-the-loop AI based lethal weapons, let us envisage a scenario in which an anti-aircraft weapons system is used by a naval vessel under attack from a squadron of manned aircraft in a theatre of war at sea in which there are no civilians present.

(B) Human-on-the-loop.

1. The *intra*-mission target selection can be undertaken without human intervention by an epistemic machine controller (a constitutive component of a LAWS), albeit in compliance with the human controllers' jointly decided *prior* target selection (for which these human controllers' are jointly (prospectively) morally responsible and, potentially, (prospectively) institutionally responsible).

²¹ Seumas Miller, *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), p. 277-278.

2. On receipt of the communication of the selected target from the epistemic machine controller, the lethal attack on this target is automatically undertaken – following on a time delay designed into the system to enable the human controller to override the system supposing new information is received regarding the target - by the kinetic machine controller without (in the generality of cases) human intervention. However, this machine controlled lethal attack is undertaken as a result of and in compliance with the human controllers' jointly decided *prior* decision to deploy human on-the-loop weaponry – a decision for which these human controllers in question (commanders) are jointly (prospectively) morally responsible and, potentially, (prospectively) institutionally responsible.
3. Due to the very large number of attackers on any given occasion of the type of combat engagement in question, there is only the practical possibility of a LAWS successfully undertaking intra mission target selection and delivery of lethal force, but not of a human operator doing so.

In this type of scenario there is the practical possibility of the human analyst (and perhaps the commander) overriding the AI based intra target selection and/or of the human operator (and perhaps the commander) overriding the automatic process from intra-mission target selection to the delivery of lethal force at any time during the engagement and, in particular, at the point in time at which the 'decision' (by the machine controller) needs to be made to commence (or not) the lethal response to what appears to be an imminent sustained enemy attack involving very large numbers of attackers. In light of this practical possibility of overriding the automatic process (immediately prior to the commencement of the process or during it), the human operator (in particular) is morally responsible and, *potentially*, institutionally responsible (perhaps jointly with the analyst (in the case of an incorrect target selection) and/or the commander), if he or she fails to override the automatic process as required (or overrides the process contrary to requirements e.g., of SOPs).

In this type of scenario there is no moral requirement for a morally informed, reasonably reliable judgement with respect to the delivery of lethal force against each of these attackers separately and in succession.

The human in-the-loop lethal weapon application outlined above manifests a moral dilemma, the correct answer to which (supposing there is always a correct answer) could vary from one military setting to another. The dilemma is that, on the one hand, without the human on-the-loop weapon, the kind of lethal attack involving a large number of manned enemy fighter aircraft in a prolonged engagement could not be defended against i.e., human controllers are unable to successfully use the weapon against this kind of attack. On the other hand, there are inherent risks in deploying a human on-the-loop weapon, notwithstanding that it is known that there are no civilians in the theatres of war in which it is to be deployed. For instance, the machine controller might mistakenly identify the incoming fighter aircraft as enemy aircraft when in fact they are the aircraft of an ally.


However, if this combination of factors is a realistic possibility in a given theatre of war then the deployment of a human-on-the-loop lethal weapon seems morally justified in which case those commanders who jointly decide to deploy and use this weapon are jointly *retrospectively* morally responsible and, potentially, *retrospectively* institutionally responsible, for the good or bad outcomes of its use (or failure to use). However, evidently such deployment ought to be planned in advance (including for the reason that it might be morally good to do so in a just war) in which case the use of this weapon in like situations might become the joint prospective *institutional* responsibility of relevant commanders and, thereby would become (also) the joint prospective *moral* responsibility of these commanders. If so, it would be an example of a prior moral problem generating an institutional solution that led, in turn, to the creation of an additional joint moral responsibility of institutional role occupants (namely, the commanders in question).

The above-described scenarios pertain to theatres of war. What of orderly jurisdictions and disorderly jurisdictions? Speaking generally, human in-the-loop lethal weapons, let alone human on-the-loop lethal weapons, ought not to be used outside theatres of war. However, if a case can be made for targeted killing of, for instance, terrorists in disorderly jurisdictions, such as the FATA, then the use of human in-the-loop lethal weapons might be justified. The justification might rely in part on the absence of alternative less harmful (including politically) means to achieve military ends, necessary military ends, (e.g., senior military leaders of terrorist organizations at the height of its power, certainty with respect to target selection, and no innocent civilians in proximity).

About the Author

Prof. Dr. Seumas Miller

Prof. Dr. Seumas Miller is Professor of Philosophy at the Australian Graduate School of Policing and Security at Charles Sturt University (Canberra), Distinguished Research Fellow at the Uehiro Centre for Practical Ethics at the University of Oxford and a Visiting Researcher at the Digital Ethics Centre at TU Delft. He is the author or coauthor of over 250 academic articles and 22 books including *The Moral Foundations of Social Institutions* (Cambridge University Press, 2010), *Shooting to Kill: The Ethics of Police and Military Use of Lethal Force* (Oxford University Press, 2016), *Institutional Corruption: A Study in Applied Philosophy* (Cambridge University Press, 2017), *Dual Use Science and Technology, Ethics and Weapons of Mass Destruction* (Springer, 2018) and *Cybersecurity, Ethics and Collective Responsibility* (with Terry Bossomaier) (Oxford University Press, 2024). He was Foundation Director of the Australian Research Council Special Research Centre in Applied Philosophy and Public Ethics (2000-2007) and Principal Investigator on a European Research Council Advanced Grant on counterterrorism ethics (2016-2021).



HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
[@hcssnl](#)

The Hague Centre for Strategic Studies