Ethics and Information Technology (2025) 27:63
https://doi.org/10.1007/s10676-025-09874-x

ORIGINAL PAPER

# Lethal autonomous weapon systems (LAWS): meaningful human Control, collective moral responsibility and institutional design

**Seumas Miller**[1,2]



**Abstract**
This article is concerned with three key ethical issues that arise from the use in military combat of Lethal Autonomous Weapons Systems (LAWS). The first issue concerns the unpredictability of LAWS in respect of the requirement of Meaningful Human Control (MHC) – as opposed to machine control - in theatres of war. It is argued that the unpredictability of (especially) 'self-learning' LAWS is not necessarily a barrier to their morally acceptable use under some restricted conditions. The second issue concerns a normative framework for ascribing moral and, potentially, legal responsibility in respect of LAWS, i.e., the so-called responsibility gap. It is argued that the notion of collective responsibility is potentially helpful here. The third issue concerns human on-the-loop LAWS. It is argued that whereas human out-of-the loop weapons should be prohibited, nevertheless, under certain restrictive conditions some human on-the-loop LAWSs (as well as human in-the-loop LAWSs) may be morally acceptable.

**Keywords** Autonomous lethal weapons · Collective moral responsibility · Meaningful human control · Responsibility gap · Institutional responsibility

## 1. Introduction

M-Q1 Predator and the M-Q9 Reaper weaponised UAVs (uninhabited aerial vehicles) or drones have been used by the US in Afghanistan, the tribal areas of Pakistan (Federally Administered Tribal Areas or FATA) and elsewhere, to kill suspected terrorists and, in the case of the Iranian General, Qasem Soleimani, a military leader of a hostile state in a third state, Iraq. While these drones were not autonomous weapons, they could be given this capability (including by means, in part, of being equipped with facial recognition technology), in which case, once programmed and activated, they could locate, identify, track and destroy human and other targets without the further intervention of a human operator. Indeed, according to a 2021 UN report (see Cetiner, 2021), in 2020 a Kagu 2 drone, having been programmed and activated, proceeded to identify, track and attack members of the Libyan National Army in Libya without human intervention. More recently, the Ukrainian military has made substantial use of, and has become heavily reliant on, semi-autonomous armed drones in the face of the Russian invasion of its territory. Indeed, facing acute manpower shortages and a lack of conventional fighter aircraft, it has adopted a novel strategy of, in effect, drone warfare (Bondar, 2025). At a more sophisticated level, fighter aircraft, such as F16s, could theoretically be operated by an AI pilot, as opposed to a human one and, if so, potentially perform a number of tasks, such as engaging in a 'dogfight', more effectively than a human could. Further developments in autonomous lethal drones include that of so-called 'swarms' of drones that operate together, and nuclear armed autonomous torpedos. Given what is at stake (in the case of even a small nuclear device) in terms of loss of human life, the threat of nuclear war etc., surely nuclear armed autonomous weapons ought to be prohibited.

There are multiple definitions of lethal autonomous weapons systems (LAWS). Here 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

✉ Seumas Miller
semiller@csu.edu.au

[1] Charles Sturt University, Bathurst, NSW (New South Wales) 2795, Australia

[2] University Offices, University of Oxford, Oxford, United Kingdom

Published online: 28 October 2025

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 (Taddeo, 2024: 173–176). The responses of such 'self-learning' systems are not entirely preprogrammed 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 (Taddeo, 2024: 198–203). Of course, non-autonomous or, at least, semi-autonomous AI based weapons system might also have a problem of unpredictability. However, the problem is likely to be more acute in the case of autonomous lethal weaponry.

Note that this unpredictability does not exclude the possibility of ascribing moral responsibility; arguably, there is not a *moral* responsibility gap (Sparrow, 2007; Steinhoff, 2013) (as opposed to a legal or regulatory responsibility gap of which more below). 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 (Miller 2016: 279).

A relevant moral principle here is the precautionary principle which has application in a number of quite different technological settings. This principle is a kind of moral meta-principle since it governs decision-making that is already governed by prior moral principles, e.g., the principle of discrimination (the principle prohibiting the deliberate targeting of non-combatants). However, there is a dearth of knowledge with respect to the potentially harmful technology in question, including the so-called 'black box' problem (Miller 2024: Ch. 4) in respect of the lack of transparency/understanding of the workings of AI to their human designers, and, as a result, compliance with these prior principles is epistemically problematic and, therefore in this context, morally problematic; hence the need for compliance with the precautionary principle. Moreover, if military leaders and others are to proceed to deploy such technology notwithstanding the risks, and in doing so ignore the precautionary principle, then they are, other things being equal, morally responsible for the bad outcomes that might ensure. For instance, a LAWS might misidentify a civilian vehicle as an enemy combatant vehicle and destroy it killing innocent persons; and the LAWS might do so without the possibility of a human controller being able to intervene at any point to prevent this adverse outcome.

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 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' is perhaps small or the robot sentry is capable only of the equivalent of small arms fire), the geographical reach, the quality of contextually relevant 'sensory' data, and the 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, (e.g., Is it to be used in the context of a genuine existential national threat such as, arguably, that being confronted by Ukraine as a result of the Russian invasion of its territory?)

Our main focus in this article is with the deployment and 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. (Santoni de Sio and van den Hoven 2018; IEEE SA 2024). 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 middle-ranking officers and above) and, secondly, engagement can be determined, in effect, by AI processes constitutive of an autonomous weapon.

Regarding the ethics or morality (these terms are used here interchangeably) of the use of LAWS, three key points are as follows. Firstly, the use of lethal force is necessarily morally significant and, as such, subject to moral constraints, e.g., in war the morally based legal principle of discrimination prohibiting the deliberate targeting of civilians.

Secondly, AI based systems, such as LAWS, do not have consciousness, cannot experience pain or pleasure, do not care about anyone or anything (including themselves), and cannot recognize moral properties, such as courage, moral innocence, moral rights and duties, moral responsibility,

compassion, mercy, justice etc. etc., or respond to moral reasons *qua moral reasons*. Therefore, they cannot act *for the sake of* moral ends or principles *qua moral ends or moral principles*. Indeed, arguably they cannot act *for the sake of an end in itself*, as opposed to as a means to some further end or in accordance with an end that has been (as a matter of contingent fact) programmed into them. Here we need to invoke a distinction between merely aiming at an end or goal and acting for the sake of that end in itself. The latter concept entails being committed to that end qua intrinsically good (or otherwise intrinsically desired) end (e.g., happiness) or qua conceptually necessary end (e.g., in the case of (mental) judgments (but not assertions), truth-aiming is (more or less) unavoidable (Miller 2015)).

Given the non-reducibility of moral concepts and properties to nonmoral ones and, specifically, to physical ones, at best computers can be programmed to comply with *non-moral proxies* for moral requirements (Arkin, 2010; Miller 2016: 279–281). Moreover, given this non-reducibility, in all likelihood, no robust, systematic, including statistical, correlation is to be found between moral properties and non-moral proxies. Certainly, none has been found to date, notwithstanding ongoing strenuous efforts (Arkin, 2010). Accordingly, AI enabled robots can only make selections based on the moral and/or legal principles and associated moral judgements of a small number of human beings (their programmers and the relevant military commanders who deploy these robots (and perhaps their political masters)). In the current Israeli-Hamas armed conflict in Gaza, for instance, the Israeli Defence Force (IDF) seems to be using AI enabled systems for lethal targeting purposes that are operating in accordance with an extremely permissive (and crude quantitative) rendering of the morally-based legal principle of proportionality. Evidently, on this rendering for any given single terrorist being targeted the lives of tens of civilians can be put at extreme risk; hence the very large number of civilian casualties in Gaza. Moreover, it is likely that AI enabled robots will never be able to find physical correlations that will enable them to mimic the *objectively correct*, *discretionary. context-dependent* moral judgements made, at least a good deal of the time, by the professionally competent, morally responsible, human commanders and combatants in actual theatres of war.

With respect to AI enabled systems that can engage in 'learning', the non-reducibility of moral properties to non-moral ones has the consequence that any such system cannot undertake moral learning *as such*. Therefore, any improvement in the levels of compliance of an AI enabled system with, for instance, the moral principle not to deliberately kill innocent persons, does not reflect an improvement in its capacity for *moral* discernment; it has no such capacity. Rather it reflects an improvement in its compliance with whatever nonmoral proxies it is sensitive to (including by virtue of its 'learned' responses to new and emerging combinations of physical entities and their properties); or it reflects its adherence to non-moral proxies of moral principles selected by their human programmer that more adequately mimic their favoured moral principles, e.g., 'Only shoot at weapons-carrying bipeds in 'light blue' (e.g., identified as the relevant short wavelength of infrared light) uniforms not in close proximity to other bipeds'.

Thirdly, war is quite unlike programming a destination into a robot-driven car with a detailed and fixed roadmap, or, for that matter, a flight path into an AI-enabled fighter aircraft. Nor is it even like playing a game such as chess, albeit it is analogous in some ways. For unlike in war, in chess there is a single, definite, unchanging, and mutually known "theatre of war" (the chessboard), a resource base that cannot reproduce itself (the chess pieces), a precise and well-defined set of rules and contexts of application, and a fixed, finite, and knowable (at least in principle) set of possible moves and countermoves (Miller 2016: 272–273).

The actual conduct of war is, or ought to be, governed by legally enshrined moral principles (e.g., so-called *jus in bello* of just war theory), notably the principles of (1) military necessity, (2) proportionality, and (3) discrimination. The application of these principles, unlike compliance with chess rules, necessarily involve moral judgements (and not merely, causal, tactical or strategic determinations). These moral principles are also quite unlike the precise and well-defined rules and contexts of application in chess (Miller 2016: Ch. 10). For instance, what counts as military necessity is typically not, and often cannot be, precisely defined. Moreover, the contexts and conditions in which they are to be applied cannot be precisely defined and, indeed, are contestable (and contestable, in part, on moral grounds). Consider, for instance, counterterrorism operations outside theatres of war yet also outside well-ordered peacetime settings, such as US drone strikes in the FATA. (The FATA is a disorderly jurisdiction in a state, Pakistan, that has not been at war with the US, notwithstanding US drone strikes on terrorists in the FATA (Regan, 2022). Further, these moral principles are unlike the rules of chess in that they necessarily undergo ongoing reinterpretation; the actual content of these principles is somewhat unstable across different settings and across time. This is in part because the entire nature of war, including the technology in use, undergoes constant change, sometimes during the course of a single war. Consider, for instance, the principle of proportionality in the context of modern urban warfare involving the potential extensive use of wide-area, explosive munitions to counter adversarial forces using buildings to conceal and protect themselves. Whether or not putting civilians at risk by the use of such munitions is proportionate in a particular

battle depends in part on what is *morally* at stake in winning the battle in question, i.e., the principles of proportionality and military necessary are conceptually interdependent (Miller 2016: Ch. 10).

The upshot of the above discussion, and like discussions, is that LAWSs ought to be subject to control by their morally sentient human operators. This is in large part because, as argued above, only human operators can understand and comply with moral principles as moral principles, including the laws of war. 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'.

Here the notion of meaningful human control (MHC) is relevant (Roff & Moyes, 2016; Santoni de Sio and van den Hoven 2018). However, the term "meaningful" is potentially ambiguous. 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). Presumably, the human control in question needs to be effective, if it is to be meaningful. Therefore, inter alia, the human controller needs to be knowledgeable with respect to the use to which the technology can be, and is being, put, the technology needs to be designed in a manner that enables the (trained) human controller to be competent in its use, the human controller needs to be technically competent to use the technology, e.g., in accordance with its standard operating procedures (SOPs) etc. Indeed, the design purpose of the technology ought to be a moral purpose (or, at least, not an immoral purpose). This is not to say that it ought to be designed so that it could never be used for an immoral purpose; that would be too high a moral standard for most, if not all, forms of weaponry. Nor is it to say that the technology itself has a purpose independent of any actual or potential human user; it is not to ascribe even humanly designed physical entities with an *inherent* teleology. However, in the case of LAWS, the following states of affairs are highly desirable, indeed morally obligatory: (i) The controller (or controllers) are morally responsible human beings (e.g., not psychopaths or otherwise significantly morally deficient); (ii) The controller (or controllers) occupy institutional roles, and are equipped with technology, both of which have been designed so as to enable them to exercise their moral judgement responsibly; (iii) The controller's (or controllers') exercise of their control results in morally correct actions (as far as is practically possible).

A key question that now arises is that of moral responsibility; the moral responsibility of human beings with respect to the control of LAWS. We begin with the concept of moral responsibility in play here.

## 2. Responsibility

There are three categories of responsibility directly relevant to our concerns in this article (Fischer, 1986; Paul et al., 1999). 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 and AI enabled robots, are able to control other entities; they are controlling agents (Russell & Norvig, 1995; Nyholm, 2023). The concept of control favoured here entails firstly, at least two entities, namely, the controlling agent (e.g., the human operator of a weapons system) and the entity over which it is exercising control (e.g., the weapon system itself), and, secondly, the causal responsibility that the controlling agent has with respect to the entity controlled. Controlling agents are persisting entities constituted by a unitary structure of properties by virtue of which they possess causal powers that enable them to *control* the actions or 'doings' of other entities, e.g., the 'doings' of weapons systems.

Entities are controlling agents in this sense (referred to here as "machine control") if they meet the following conditions. First, controlling agents cause their controlled entities to perform certain actions or otherwise do certain things ('doings'), e.g., the controlled weapon fires at a particular target. Second, controlling agents can access and process data (e.g., in accordance with algorithms) in a manner that enables this data to 'inform' the exercise of their causal kinetic powers, e.g., facial recognition technology enables the identification of a potential target for the weapon of a LAWS. Third, controlling agents can exercise their causal powers in a manner that involves them selecting the particular action or 'doing' to be done by the controlled entity from an array of options, depending on prevailing circumstances, e.g., the weapon fires if the target is a biped carrying a weapon but only if there are no non-weapon-carrying bipeds in close proximity. Fourth, controlling agents make use of data feedback loops that enable them to monitor, adjust and correct the 'doings' of the controlled entity, e.g., a weapon system that responds to defensive moves on the part of its target by virtue of its AI enabled controlling agent. Fifth, with respect to certain key tasks, e.g., selecting and attacking targets, the controlling agent has the ability to act or otherwise do things independently (to some non-negligible degree) of other controlling agents, e.g., LAWS can operate independently of human operators. That is, a controlling agent cannot be under the *complete* (or even near-complete)

control of some other controlling agent on pain of ceasing to be a controlling agent.

Clearly this notion of control, while presumably more or less definitive of, for instance, machine control, does not define, but rather presupposes, simple causation. Thus, a struck billiard ball that collides into a second billiard ball causing the latter to move forward is not thereby exercising control over the second billiard ball. However, machine control (thus defined) 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 that involves, at least, the ability to (non-randomly) do otherwise than he or she actually does – something not true, presumably, of machine control) and the controller's responsiveness in the exercise of its control to reasons (qua reasons), including *moral reasons* (qua moral reasons). Note also that the laws/regulations (e.g., International Humanitarian Law (IHL), Law of Armed Conflict (LOA), 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 equipment (e.g., weapons), and so on are (at least to a considerable degree) under human control (or, at least, ought to be). Moreover, 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/regulation, military missions, ROEs, SOPs 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. (In this section the possibility of ascribing responsibility to institutions per se is for the most part ignored (Miller 2010: Ch. 4). But see section 5.) As with moral responsibility, the bearers of institutional responsibility are human beings, but not volcanoes or AI enabled robots. Note that some institutional responsibilities involve the exercise of institutional authority over other persons, including other role occupants. However, the exercise of institutional authority by superiors over their subordinates does not necessarily extinguish the institutional, let alone, moral, responsibility of subordinates (notwithstanding the prominent legal notion of command responsibility in military organisations for many of the actions of subordinates).

Note also that since the occupants of institutional roles are human beings, institutional role occupants possess many prior moral rights and duties that they possess by virtue of being human beings, e.g., the natural (as opposed to institutional) moral right of personal self-defence. That said, some institutional rights and duties may entail the suspension or curtailment of such prior moral rights and duties, e.g., the prior natural moral right of freedom of movement, expression etc. is curtailed once he or she enlists in the armed forces.

Moreover, institutional arrangements assign moral responsibilities to human beings that they did not previously have; in the case of the institutional role of combatant, for example, the responsibility to wage war and, in so doing, put his or her life at risk. The moral basis for this latter institutional arrangement appears to be collective self-defence (including of the collective's sovereign territory). Once institutions and their constitutive roles have been established on some adequate moral basis then those who undertake these roles necessarily put themselves under institutional and moral obligations of various kinds—obligations that attach to, and are in part constitutive of, those roles (Miller 2010: Ch. 2).

However, 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 to shoot dead enemy combatants. 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. 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 (Miller 2025). 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) (Miller 2016: Ch. 10). 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. They cannot be held criminally liable since they do not have *mental* states, notably intentions (specifically, mens rea) but rather only have *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 and rationally connected to other mental states,

such as beliefs, desires and emotions, and are such that the agents possessed of them are, or can become, conscious of them; but AI enabled robots do not have this kind of web of mental states constitutive of human rationality (hence the 'black box' problem) 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'.

It is important to note that institutional responsibilities can be attached by design to institutional roles (and new institutional roles designed from scratch) in a manner that not only tracks prior moral responsibilities but to some extent creates new individual moral responsibilities. Moreover, once these institutional roles are created, their role occupants can be held institutionally 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. The significance of this point for our concerns here is that so-called responsibility gaps can potentially be closed by recourse to this process of redesigning institutional roles and, thereby, generating institutional and moral responsibilities constitutive of these roles, notably *collective* institutional and moral responsibilities, that may not have existed prior to this process.

## 3. Collective responsibility

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 (Miller 2006). On this view, collective responsibility is responsibility arising from joint actions and omissions. Roughly speaking, a joint action can be understood thus: Two or more individuals perform a joint action (Miller 1992) if each of them intentionally performs an individual action (or omission) but does so with the (true) belief that in so doing they will jointly realise an end which each of them has and has interdependently with the other (a collective end in our parlance). On this view of collective responsibility as joint responsibility, collective responsibility is ascribed to individual human beings only, albeit jointly. Each member of the group is individually morally responsible for his or her contributory action and also for the outcome of the set of actions. However, each is individually responsible for that outcome, *jointly with the others*; hence the conception is relational in character.

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 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.

Here we need to distinguish between kinetic action, e.g., the firing of the weaponised drone, and non-kinetic *epistemic* action (from the Ancient Greek word, episteme, meaning knowledge). Epistemic actions have a constitutive epistemic end such as knowledge or belief, e.g., knowledge that the man in the scenario is (or is not) laying an IED. Military intelligence is essentially epistemic activity, albeit in the service of kinetic activity (e.g., 'killing people and breaking things').

In relation to moral and institutional responsibility, we need to distinguish retrospective from prospective responsibility. Retrospective responsibility is backward looking and pertains to an action, omission or outcome that has actually taken place. The exercise of retrospective responsibility frequently results in praise or blame for these past actions, omissions or outcomes. However, the concept of being responsible (retrospectively, let alone prospectively) is not the same concept as being blameworthy or praiseworthy (although the latter concepts are presupposed by the concept of being responsible) (Miller 2006). Nor is the concept of retrospective responsibility to be identified with the concept of liability.

Prospective responsibility is forward looking and pertains to actions, omissions or outcomes that are yet to take place and might never take place. The concepts of prospective and retrospective responsibility go hand in glove in that

if one had the prospective responsibility to perform some action and failed to perform it, then one would be retrospectively responsible for failing to perform it (and if one in fact performed it then one would be retrospectively responsible for doing so). Relatedly, prospective responsibility for certain actions frequently involves liability for those actions (including for failure to perform them). However, prospective responsibility and liability (let alone accountability) are not the same concepts.

Our concern in this article is with an important species of prospective responsibility, namely that which attaches to institutional role occupants. The above-mentioned operator, for instance, has the prospective role responsibility (once authorised) to see to it that the weaponised drone hits its target. Indeed, given the moral significance of this action, the operator is also morally responsible for this action. The members of the drone crew who analyse the visual imagery have the joint institutional responsibility to see to it that they correctly determine that the man is (or is not) laying an IED. Likewise, given the moral significance of this joint epistemic action, they are also jointly morally responsible for this determination.

Moreover, the soldier on the ground and the members of the drone crew (including the commander) are jointly morally responsible for killing the terrorist (and, presumably, praiseworthy for doing so, given that they have saved innocent lives in doing so). However, it is a further question whether they are jointly *institutionally* responsible for killing the terrorist. That would depend on the institutional arrangements in place. Perhaps, for instance, the commander *alone* is institutionally responsible for the killing of the terrorist on the institutionally-based grounds that the commander accepted the determination of the drone crew with respect to the visual imagery and the commander commanded the operator to fire. At any rate, an important question that arises here concerns the adequacy of such institutional arrangements. More generally, what ought to be the institutional arrangements, and associated individual and joint prospective institutional responsibilities, with respect to LAWS, including in the design phase and deployment phases, as well as in their use? However, these questions presuppose an understanding of the nature of the human-machine interaction in LAWS.

## 4. Human-Machine-Interaction

We need to get some clarity on the general nature of human-machine interaction in respect of LAWS. A standard distinction is between so-called '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 in the required sense. In the case of human-in-the-loop weapons the delivery of lethal force (for example by a predator drone), cannot be done without the decision to do so by the human operator. In the case of human on-the-loop weapons, the delivery of lethal force can be done without the decision to do so by the human operator; however, the human operator can override the weapon system's triggering mechanism. In the case of human out-of-the-loop weapons, the human operator cannot override the weapon system's triggering mechanism; so once the weapon system is programmed and activated there is, and cannot be, any further human intervention. Naturally, an autonomous weapon, i.e., a 'human-out-of-the loop' weapon, could, nevertheless, be designed so that it could in fact also function as a 'human-in-the-loop' or 'human-on-the-loop' weapon, supposing their human commanders or operators decided to activate one or other of these functions prior to engagement. However, it is very difficult to see what reasons could be given for designing a human out-of-the-loop weapon, as opposed to a human on-the-loop weapon (or, of course, a human in-the-loop weapon). At any rate, with respect to so-called autonomous weapons, it is important to bear in mind the distinction between autonomy (e.g., human in, on or out of, the loop) at the level of design and autonomy at the level of actual use.

Let us get further clarity on these issues by initially considering a relevantly similar human-human interaction; specifically, a simplified version of our above-described weaponised drone scenario. This version involves just two human persons, Analyst and Operator. In this scenario Analyst analyses the visual imagery and makes the determination that the suspected terrorist is placing an IED, and Operator fires the weapon which kills the suspected terrorist, relying on Analyst's determination to do so.

In this scenario there is a joint action that consists in part of, firstly, Analyst making a determination and communicating it to Operator (having as an individual end that Operator knows that the suspected terrorist is in fact placing an IED), and, secondly, Operator firing the weapon (having as an individual end to kill the target). So the joint action consists in part in these two individual actions (the first of which is an epistemic action, the second a kinetic action). However, it is more than this since, firstly, these two actions are interdependent: Operator will only fire the weapon if Analyst communicates Analyst's determination that the target is in fact laying an IED; and Analyst would not communicate this unless Analyst knew (or reasonably believed) that Operator would fire the weapon at the target which Analyst had determined was laying an IED. Moreover, these two individual actions are interdependent by virtue of an end *shared* by Analyst and Operator, namely, that the target be killed if and only if it is known that he is laying an IED. Further, the

possession by each of the two of this shared end is interdependent (at least, if it is a collective end and the action is a joint action (Miller 1992), i.e., neither would have this end if the other did not also have it, and if each did not know that the other had this end[1] (presumably because, at least in part (and notwithstanding any hierarchy of institutional roles), neither could achieve it without the assistance of the other and the other would not assist unless he or she also had this end).

Importantly, in our example, it might be thought that Operator could be an AI enabled robot, as could Analyst. If so, then the drone in the scenario would presumably be a species of LAWS.

Moreover, if so, then it would seem that robots can perform joint actions or, at least, joint 'doings'. Here we must note that for reasons mentioned above, notions of human control (and, relatedly, human autonomy) are out of place here; rather the notion of machine control is in play. However, AI enabled robots do not have human control but only machine control and, as a result, they cannot interact jointly in other than a purely instrumental manner (or, at least, engage in a causal process that mimics an instrumental (and, therefore, teleological) relation).

One consequence of this might seem to be that the only form of joint machine control is one in which each controlling agent acts individually in terms of its prior individual 'ends' (which could include shared individual 'ends' (even if not shared *interdependent* ends in the relevant sense -see below) such as is the case with, for instance, a weapon that requires both operators to 'pull the trigger' if it is to fire. The only alternatives to this form of instrumentally based joint control might be single i.e., non-joint, sources of control (or an absence of control). Moreover, in order for joint machine control to be genuinely joint it would need to be the case that the shared 'ends' are interdependent, i.e., each would have as an end that it 'pull the trigger' if and only if the other does. This might lead to the problem of an absence of control or, at least, of deadlock, given that each could not pull the trigger unless the other had already done resulting in a 'chicken and egg' problem with respect to cause and effect; if x causes y then y cannot cause x (and vice-versa). The resolution of this problem might seem to have to consist in one or other of the AI based robots exercising control over the other, i.e., determining whether the trigger is to be pulled (perhaps by way of a one-way veto over the other robot). If so, the robots would not be engaged in a genuinely joint 'doing, let alone a joint action. Note that by contrast human beings can, at least in principle, resolve such deadlocks without recourse to a single source of control (and without abandoning the attempt to control altogether). They can do so by virtue of their possession of freewill and, especially, human autonomy, and associated relations of equality and mutual respect, which inter alia enables morally informed negotiations.

Reverting to our earlier scenario, if Operator and Analyst are AI enabled robots who can bring it about, apparently jointly, that the suspected terrorist is killed then Operator and Analyst each perform their pre-programmed tasks, and each does so in the service of an 'end' that each individually has. Moreover, the 'doing' of each in the service of their respective 'ends' involves dependence. Operator would not fire at the terrorist laying an IED, if Analyst had not identified him as laying an IED. However, this dependency would seemingly only be one way since Analyst could determine the identity of the potential terrorist independently of whether Operator was to shoot (or not shoot). Moreover, in this scenario, neither Analyst nor Operator has the *shared*, *interdependent* (i.e., collective) end described above in the counterpart human-human interaction scenario, namely, that the target will be killed if and only if it is known that he is laying an IED. Rather each has a different individual 'end' and these two different 'ends' are causally connected (i.e., there is seemingly only the one way dependence of Operator's dependence on Analyst). That is, it is (again) not a genuine case of a joint 'doing, let alone a joint action.

Accordingly, it seems that it is not only the case that robots cannot perform actions and, therefore, not joint actions, but also that they might not be able to perform genuinely joint 'doings'. However, two machine controllers can have shared 'ends' (and one-way dependence with respect to these 'ends') and the resulting combinations of 'doings' (quasi joint 'doings') mimic (to some extent) some joint actions that human perform. Moreover, at least in principle, the repetitive performance of 'doings' as components of quasi joint 'doings' might constitute a role of sorts in an institution, e.g., that of Operator of a weapons system, even if not a full-blown institutional role.

However, in occupying such a role in an institution the robots would not be choosing their 'ends'; rather, as mentioned above, these would be programmed into them. Further, as we saw above, institutional roles are defined in part by institutional rights and duties, many of which are also moral rights and duties, and even AI enabled robots are not moral agents capable of recognising, and acting for the sake of, moral considerations. So AI enabled robots do not occupy institutional roles in the full sense of that term.

Let us now turn directly to actual applications of LAWSs and, in particular, applications involving human-machine interaction these being, evidently, the only morally acceptable applications of LAWSs.

[1] Indeed, there would be mutual knowledge to this effect. A and B mutually know that p if A knows that p, B knows that p, A knows that B knows that p, B knows that A knows that p, and so on. (Smith, 1982).

## 5. LAWSs' applications: integrating technological, institutional and moral dimensions

Firstly, in relation to LAWS technology, we need to invoke the above-mentioned 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 (in our favoured sense of 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) (Miller 2016: Ch. 10.2). Moreover, we also assume that if a lethal human-in-the-loop weapons 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 Sect. 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 LAWSs, we need to invoke the above-mentioned distinctions between analysts (target identification and selection), operators (firing weapons) and the commanders thereof. We also need to invoke distinctions between orderly jurisdictions in peacetime, theatres of war, and disorderly jurisdictions without effective law enforcement that are experiencing ongoing, serious, armed conflict e.g., the FATA (Miller 2009: Ch. 5; Miller 2016: 265) and between using lethal force against combatants in close proximity to innocent civilians and not doing so. We assume that it would be morally unacceptable to use LAWSs in orderly jurisdictions in peacetime, i.e., for law enforcement purposes.

Thirdly, we need to invoke a distinction between decision-making made prior to a combat mission and decision-making (or, in the case of LAWSs, 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 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 G is underway, then the identification of a person as being a target because he is an enemy combatant in area G would be an example of *intra* mission target selection. As noted above, 'self-learning' LAWSs ought not necessarily to be prohibited because, for instance, they cannot meet MHC requirements by virtue of their unpredictability and the resulting unintentional, unjust harm that they might cause. This argument evidently fails given that the likely degree of unpredictability may well be low if, for instance, the LAWS in question is prevented from 'learning' and updating its functioning while undertaking a mission, (intra mission 'learning' and updating), its firepower is small, the geographical range of the weapon is short, there are no civilians in proximity and the militarily 'necessary' end is of very great strategic and, therefore (in a just war) moral importance.

The use of a human-in-the-loop, even if AI enabled, lethal weapon in a theatre of war in compliance with relevant IHL, LOAC and the related ius 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 areas 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, 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 and a defensive weapon. Such a human in-the-loop LAWS might be notionally characterised in terms of the two conditions (1) and (2) under the heading Human-in-the-loop (Miller 2016: 277–278) (although there are a variety of possible alternative institutional arrangements to the ones described 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 in part of the target selection determinations of the AI enabled data collection and analysis processes of LAWSs, e.g., facial recognition software;

(2) In accordance with their respective institutional roles, human controllers (the operator and the commander) jointly initiate a LAWS' lethal attack and in doing so rely on the intra-mission target selection provided by the analyst (on the basis in part of the target determination of the LAWS) - and accepted by the commander.

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. 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 very likely not only 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 epistemic 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). Importantly, the analysts need to be able to resist any inclination to become over-reliant on AI-based intra target selections.

Moreover, the operators involved are also (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 kinetic actions (and, potentially, for 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 consult their commander.

In light of the above discussion, we can conclude that analysts, operators and their commanders are jointly (prospectively) morally responsible, and, potentially, jointly (prospectively) institutionally responsible , 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. If so, then there could be grounds for ascribing regulatory and even 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) (Miller 1995).

The institution or institutional unit to be held criminally liable for some jointly produced (including by joint omission) bad outcome (including outcomes to which many members made very small contributions and/or performed individual contributory actions that would not normally be regarded *in themselves* as crimes) could be circumscribed by recourse to the individual members of that institution or unit who were jointly (prospectively and retrospectively) *culpably* morally and institutionally responsible for the bad outcome in question. The scope of the unit held criminally liable might (potentially) include members who were responsible only for failing to take steps to prevent others from contributing to the bad outcome or for failing to report those who had already contributed to the bad outcome. However, the punishment would depend in large part on the seriousness of the outcome of the joint action or omission, e.g., the quantum of lives lost as a result of an out of human control LAWS. Moreover, the punishment could be applied differentially to some extent, depending on the contribution of each to the bad outcome, e.g., other things being equal commanders would be liable for a greater punishment than their subordinates.

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 outcome and/or the operator (and perhaps the commander) has the ability to override the weapons system. The potentially justified ability to override, in our favoured sense, is a species of *human meta control* such that : (i) The operator has the ability to shut down the weapon; (ii) The operator has 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.

Moreover, the human operator is, perhaps jointly with others (such as his or her 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 assumptions, the salient ones of which are outlined below (Miller 2016: 277–278).Such human on-the-loop LAWSs might be notionally characterised (under the heading Human-on-the-loop) in terms of the conditions, (1), (2) and either condition (3)(i) or condition (3)(ii) or condition (3)(iii). Accordingly, there are three species of human on-the-loop LAWs identified.

The first of three species has the following three characterising conditions.

(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).

(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 to enable the human controller to override the system - 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) (i) On each and every occasion of intra mission target selection and delivery of lethal force, there is, firstly, the practicable possibility of the human analyst or commander overriding the target selection, and, secondly the practicable possibility of the commander or human operator overriding the automatic process from target selection to the delivery of lethal force. Moreover, the human analyst, the commander and the human operator have sufficient time and sufficient information to make their respective morally informed, reasonably reliable judgements whether or not to override LAWS' target selection or its deliver of lethal force.

The second species of human on-the-loop LAWS is characterised by conditions (1) and (2) above but also by condition (3)(ii) (as opposed to condition (3)(i)).

(3) (ii) Due to the very large number of attackers on any given occasion of the type of combat engagement in question, there is only the practicable possibility of a LAWS *successfully* undertaking intra mission target selection and delivery of lethal force, but not of a human operator doing so. Nevertheless, there is the practical possibility of the human analyst or commander overriding the automatic process of intra-mission target selection or of the commander or human operator overriding the automatic process from intra 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 (and perhaps the commander) is morally responsible and, *potentiall*y, institutionally responsible, if he or she fails to override the automatic process as required (or overrides the process contrary to requirements e.g., of SOPs). 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 third species of human in-the-loop LAWS is characterised by conditions (1) and (2) above but also by condition (3)(iii) (as opposed to condition (3)(i) or condition (3)(ii)).

(3) (iii) There is a single attacker but, nevertheless, there is only the practicable possibility of a LAWS *successfully* undertaking intra mission target selection and delivery of lethal force, but not of a human analyst and operator (or, therefore, commander) doing so. There is the physical possibility of the human analyst or the human operator (with or without the permission of the commander) overriding the automatic process from intra-mission target selection to the delivery of lethal force (including, in the case of the human analyst, by overriding the intra target selection) but in neither case, given the time constraints, is it practicable to do so on the basis of morally informed, reasonably reliable judgements. There is a moral requirement for morally informed, reasonably reliable judgements with respect to target selection and the delivery of lethal force against this single attacker, given the likelihood of significant (but not massive) collateral harm to civilians in close proximity. However, if the LAWS' intra-mission target selection is in fact correct (even though this is unlikely), and if the attacker is not destroyed, then the harm that will be done by the attacker, in terms of innocent civilian life, will be massive.

The human in-the-loop lethal weapon application outlined in (3)(ii) manifests a moral dilemma the correct answer to which (supposing there is a correct answer) could vary from one military setting to another. The dilemma is that without the human on-the-loop weapon the kind of lethal attack involving, let us assume, 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 could be morally justified (depending on likely outcomes, what is morally at stake and so on) 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 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 human in-the-loop lethal weapon application outlined in (3)(iii) also manifests a moral dilemma, albeit in a more acute form, the correct answer to which (supposing there is one) could vary from one military setting to another. However, if this combination of factors is a realistic possibility in a given theatre of war then the deployment of a human-in-the-loop lethal weapon could be morally justified (depending on likely outcomes, what is morally at stake and so on) 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, it might also be the case that such deployment ought to be planned in advance (including

for the reason that it would be morally good to do so) 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).

A scenario illustrating (3)(ii) might be an anti-aircraft weapons system being 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. A scenario illustrating (3)(iii) might include ones involving a kamikaze pilot or suicide bomber set to deliver a massively destructive bomb on a city.

The above-described scenarios all 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 Islamic State at the height of its power, certainty with respect to target selection, and no innocent civilians in proximity.

**Author contributions** Single authorship only, i.e., Seumas Miller.

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing interests** The authors declare no competing interests.



## References

Arkin, A. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics*, *9*, 332–341.

Bondar, K. (2025). *Ukraine's future vision and current capabilities for waging AI-enabled autonomous warfare*. Centre for Strategic and International Studies.

Cetiner, Y. (2021). Turkish Kagu-6 Carries out First Autonomous Drone Attack, UN Report Says *Overt Defense* https://www.overtdefense.com/2021/06/02/the-turkish-kargu-2-carries-out-the-first-autonomous-drone-attack-un-report-says/ (Accessed 8/3/2025).

de Santoni, F., & van den Hoven, J. (2018). Meaningful human control over autonomous weapons. *Frontiers in Robots and AI*, *5*, 15.

Fischer, J. M. (Ed.). (1986). *Moral responsibility*. Cornell University Press.

IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems. (2024). *A framework for human decision making through the lifecycle of autonomous and intelligent systems in defense applications*. IEEE SA.

Miller, S. (1992). Joint Action. *Philosophical Papers, 21*(3), 275–299.

Miller, S. (1995). Corporate Crime, the Excesses of the 80's and Collective Responsibility. *Australian Journal of Corporate Law, 5*(2), 39–51.

Miller, S. (2006). Collective Moral Responsibility: An Individualist Account. *Midwest Studies in Philosophy, 30*, 176–193.

Miller, S. (2009). *Terrorism and Counter-Terrorism: Ethics and Liberal Democracy*. Blackwell.

Miller, S. (2010). *The Moral Foundations of Social Institutions*. Cambridge University Press.

Miller, S. (2015). Joint Epistemic Action and Collective Responsibility. *Social Epistemology, 29*(3), 280–302.

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

Miller, S. (2025). Robots, Institutional Roles and Joint Action. *Ethics and Informationa Technology, 27*(1), 1–11.

Miller, S., & Bossomaier, T. (2024). *Cybersecurity, Ethics and Collective Responsibility*. Oxford University Press.

Nyholm, S. (2023). *AI and Ethics*. 3, 1229–1239.

Paul, E. F., Author, F., & Paul, J. (Eds.). (1999). *Responsibility*. Cambridge University Press.

Regan, M. (2022). *Drone strike: Analyzing the impacts of targeted killing*. Palgrave-Macmillan.

Roff, H., & Moyes, R. (2016). 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) Geneva, 11–15 April. https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Prentice Hall.

Smith, N. (Ed.). (1982). *Mutual knowledge*. Academic.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*, 63–77.

Steinhoff, U. (2013). Killing them safely: Extreme asymmetry and its discontents. In B. J. Strawser (Ed.), *Killing by remote control: The ethics of an unmanned military*. Oxford University Press.

Taddeo, M. (2024). *The ethics of artificial intelligence in defence*. Oxford University Press.