Philosophy & Technology (2022) 35: 78
https://doi.org/10.1007/s13347-022-00571-x

RESEARCH ARTICLE

Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—a Moral Gambit

Mariarosaria Taddeo[1,2] · Alexander Blanchard[2]

Received: 8 May 2022 / Accepted: 27 July 2022 / Published online: 5 August 2022

Abstract
In this article, we focus on the attribution of moral responsibility for the actions of autonomous weapons systems (AWS). To do so, we suggest that the responsibility gap can be closed if human agents can take *meaningful moral responsibility* for the actions of AWS. This is a moral responsibility attributed to individuals in a justified and fair way and which is accepted by individuals as an assessment of their own moral character. We argue that, given the unpredictability of AWS, meaningful moral responsibly can only be discharged by human agents who are willing to take a *moral gambit*: they decide to design/develop/deploy AWS despite the uncertainty about the effects an AWS may produce, hoping that unintended and unwanted or unforeseen outcomes may never occurs, but also accepting to be held responsible if such outcomes will occur. We argue that, while a moral gambit is permissible for the use of non-lethal AWS, this is not the case for the actions of *lethal* autonomous weapon systems.

Keywords Artificial intelligence · Autonomous weapons systems · Lethal autonomous weapons systems · Meaningful moral responsibility · Moral gambit · Moral responsibility · Responsibility gap

1 Introduction

Only humans are morally responsible for the actions of autonomous weapons systems (AWS). This is because intentions, plans, rights, and duties, praise or punishment can only be attributed in a meaningful way to humans. Placing this moral responsibility on AWS, or in general on artificially intelligent (AI) systems, would entail misplacing:

✉ Mariarosaria Taddeo
mariarosaria.taddeo@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, Oxford, UK

[2] The Alan Turing Institute, London, UK

> causal accountability and legal liability regarding their mistakes and misuses. Robots could be blamed and punished instead of humans. And irresponsible people would dismiss the need for care in the engineering, marketing and use of robots (Floridi & Taddeo, 2018, 309).

Over the past years, consensus on placing moral responsibilities for the actions of AWS on human agents has grown. By now, this position is uncontroversial. For example, the UN Group of Governmental Experts on emerging technologies in the area of Lethal Autonomous Weapons Systems of the Convention on Certain Conventional Weapons (CCW) identifies human responsibility as a key principle for the (possible) use of lethal autonomous weapons systems (LAWS), stating that:

> Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire lifecycle of the weapon system (UN GGE CCW, 2019).

Human responsibility is mentioned explicitly in the ethical principles for the use of AI in the Recommendations on the Ethical Use of Artificial Intelligence by the US Department of Defense, issued by the US Defense Innovation Board (DIB), whose first principle states that:

> Human beings should exercise appropriate levels of judgment and remain responsible for development, deployment, use and outcomes of DoD AI systems (DIB, 2020a, 8).

Similarly, in a recent report to European Parliament, the Committee for Legal Affairs stated that:

> […] autonomous decision-making should not absolve humans from responsibility, and that people must always have ultimate responsibility for decision-making processes so that the human responsible for the decision can be identified (Lebreton, 2021).

The consensus on this point is important, for it avoids the risk of anthropomorphising AWS, and AI systems more broadly, and saves time from engaging with literature discussing whether AWS can be morally responsible for their own actions. Insofar as these systems do not have intentions, and no understanding of the blame or praise that may follow their actions, AWS do not bear moral responsibility for their own actions.

As shown in the rest of this article, while the focus on human agents points the debate in the right direction, it remains problematic to ascribe moral responsibilities to humans for the actions of AWS in a *meaningful* way (more on this presently). Nonetheless, ascribing this moral responsibility to individuals—as opposed to institutions or legal entities—is a necessary condition for the deployment of AWS. As it was stressed in the Nuremberg trials:

> Crimes against international law are committed by men, not by abstract entities, and only by punishing individuals who commit such crimes can the provi-

> sions of international law be enforced (International Military Tribunal (1947), 221).

In the age of autonomous warfare, AWS may perform immoral actions, but it is only by holding the individuals who design, develop and deploy them morally responsible that the morality of warfare can be upheld.

In the rest of this article, in Sect. 2 and its subsections, we analyse the main contributions to the debate on the moral responsibility for AI systems in general. In Sect. 3 and its subsections, we focus on specific approaches to ascribing moral responsibility for AWS and consider the limits as well as the points of strength of these contributions. In Sect. 4, we offer our own contribution to the debate by focusing on the concepts of *meaningful moral responsibility* and *moral gambit*. Following the analysis provided in this section, in Sect. 5, we provide eight recommendations addressing specifically how defence institutions can overcome the responsibility gap for the use of non-lethal AWS. In Sect. 6, we conclude the article.

Before beginning our analysis, three clarifications are necessary. First, in this article, we will focus on AWS, understood as:

> an artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal without the direct intervention of another agent, and may be endowed with some abilities for changing its transition rules to perform successfully in a changing environment, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object or a living being) and to this end is able to identify, select and attack the target without the intervention of another agent is an AWS. Once deployed, AWS can be operated with or without some forms of human control (in, on or out the loop) (Taddeo & Blanchard, 2021)

This is a value-neutral definition, so it has no other bearing on our analysis than to identify the set of AWS to which it applies. It is worth noting that this definition of AWS differs from many others insofar as it considers explicitly adapting capabilities as a key feature of AWS. It is this characteristic which drives much of the discussion in this article around the attribution of moral responsibility for the actions of AWS.

This takes us to our second clarification. Following our definition of AWS, lethal and non-lethal AWS are distinguished on the basis of the *purpose* of their use and not the effect of use. A lethal AWS is used with the goal of exerting lethal force—i.e. resulting in death or permanent injury—against human beings. A non-lethal AWS is used with the purpose, as with non-lethal weapons generally, of incapacitating human beings "without causing death or permanent injury" (Davison, 2009, 1). Thus, following this distinction a non-lethal AWS may also be, for example, a system used with the purpose of material destruction or disabling equipment. As we argue in the rest of this article, the outcome of AWS is predictable only to an extent. Thus, it is conceivable that an AWS used for non-lethal purposes may produce lethal effects (Coleman, 2015; Enemark, 2008a, 2008b; Kaurin, 2010, 2015; Heyns, 2016a, 2016b). This is because there is "a potential disconnect between the intention behind the use of a weapon and the consequences thereof" (Enemark, 2008b, 201).

However, one can focus on the purpose of deployment (and effective conditions of deployment) as a way of distinguishing lethal from non-lethal AWS. This focus is sufficient to support the analysis developed in the rest of the article, which aims at dealing precisely with the range of unpredictable outcomes of AWS. In the rest of this article, part of the analysis of the moral responsibility of AWS applies squarely to both the cases of LAWS and non-lethal AWS. We will distinguish between "non-lethal AWS" and LAWS whenever our analysis of the case of LAWS leads to different outcomes than the one concerning non-lethal AWS.

Our third clarification addresses the nature of the responsibility on which we will focus. Our goal is to understand how a human agent can take *meaningful moral* responsibility for the actions of AWS. This implies that moral responsibility is not ascribed nominally to human agents, for example because of their role or ranking, but bears in a justified and fair way on those who have played a key role in the realisation of the effects of using AWS. This also implies that a human agent accepts this moral responsibility, and the praise or blame that come with it, as an individual (in a personal sense) and not as a member or representative of a defence organisation or of a professional body. The attribution of meaningful moral responsibility may underpin legal processes to attribute legal responsibility, define accountability and liability. However, while related to legal responsibility, moral responsibility differs from it. For example, as we shall see in Sect. 2, the attribution of moral responsibility and the subsequent blame/praise requires causal and intentional connection between an action and effect. This is not necessarily the case when considering legal responsibility. Consider, for example, the concept of "faultless responsibility", according to which punishment can be ascribed even the intention to determine a given effect cannot be determined. Our work focuses only on moral responsibility, and while it may provide the conceptual ground for attributing legal responsibility, it is not concerned with the latter.

With the conceptual space of our analysis outlined, we can now delve into the literature focusing on attributing moral responsibility for AI systems.

2 Moral Responsibility for AI Systems

The debate on the moral responsibilities for the actions of AWS hinges on the wider discussion of the moral responsibilities for the actions of AI systems. As mentioned in Sect. 1, while there is growing consensus that these responsibilities rest with human agents, ascribing them correctly has proven to be problematic. This is because, under classic ethical approaches, for the attribution of moral responsibility to be justified and fair, agents must have a specific relation to their own actions, and their consequences. This relation must satisfy four conditions at the same time:

- Intentionality condition: the intentions of the agent to achieve a given effect (Branscombe et al., 1996; Kant & Borken, 2019; Khoury, 2018).
- The causality condition: there has to be a causal connection between the decision/action of the agent and their effects (Fischer & Ravizza, 2000; Sartorio, 2007; Shoemaker, 2017).

- Consequence condition: the agent has to have an understanding of the effects of the decision/action, as well as of their moral value and the consequent blame/praise (Bentham, 1789; Wallace, 1998; Neil Levy, 2008; Kelly, 2012).
- Choice condition[1]: the agent has some degree of freedom that allows him/her to choose among different patterns of actions (Strawson, 1962; 1962; Watson, 1975; Nelkin, 2011).

The first three conditions are problematic to satisfy when considering AI systems. Problems with meeting these conditions underpin the "responsibility gap" (Floridi, 2012; Matthias, 2004). Let us consider the intentionality condition first. Moral responsibility is often ascribed with the aim of distributing blame/praise to individuals for their morally evil/good actions. For this, intentionality of the agent is a key element to ensure that the allocation of moral responsibility is justified:

> It would be counterproductive to attribute responsibility, and hence allocate blame or praise, punishments or rewards, if the agents' actions were not intentional, because such attribution would then be *arbitrary* and *indistinguishable* from a mere random allocation, which would defy the purpose of blame or praise, punishments or rewards, […]. (Floridi, 2016, 4, emphasis added).

Under a classic ethical approach, lack of intentionality undermines the allocation of moral responsibility and of the blame/praise linked to it, even when the causal chain of events leading to a given outcome is clear. When considering AI systems, their behaviour may not result directly from the intentions of the individual designers, developers or deployers. This can be the case for two reasons: distributed actions and lack of predictability. AI systems can perform morally loaded actions which stem from a number of morally neutral actions—i.e. individual actions performed by humans or other artificial agents and which taken individually do not lead to specific morally evil/good outcomes (Floridi, 2016). We can imagine a network of agents involved in the design, development and use of an AI system, each making morally neutral decisions, which once coordinated at the network level lead to an morally evil/good action. This is what (Floridi, 2012) calls distributed morality. While the entire network can be held morally responsible for these actions, attributing blame or praise individually to all, or any of, the agents in the network would be unjustified insofar as individual actions per se do not lead to any morally loaded outcome even if, once coordinated at network level, these actions cause the final outcome. Distributed morality is not unique to AI systems, but it is particularly relevant in the AI domain, where the network of involved agents and the subsequent fragmentation of tasks among them are particularly widespread, making it problematic to identify intentionality for actions and attribute moral responsibility accordingly. At the same

[1] In the rest of this article, we will not focus on the choice condition, as in the relevant literature this refers to metaphysical understanding of determinism and freedom. The condition focuses on whether humans are or not fully determined and, thus, can choose among alternative pattern of actions. Whether this condition is met or not is independent from the case of AI systems and more related to the metaphysical view one has.

time, once deployed, some AI systems may develop autonomously new behaviours that were never intended by the human agents designing, developing and deploying these systems and for which they cannot be held morally responsible, given this lack of intentionality (more on this in Sect. 2.1).

Respecting the causality condition when considering AI systems is also problematic insofar as intended actions at design, development and deployment stages may cause an unintended outcome. To put it more precisely:

> intentionality is not closed under causal implication, […]. In the *direct* case of non-closure, it is not the case that, if Alice means to cause *a*, and *a* causes *b*, it follows that Alice means to cause *b*. (Floridi, 2016, 4).

This is particularly relevant when considering AI systems that may develop unforeseen and unintended behaviour (i.e. lack of predictability). This lack of predictability makes it problematic to meet the consequence condition. Insofar as the outcomes of the system that one deploys cannot be foreseen, it is not feasible for the agent to consider all possible consequences of this action and their moral value.

Lack of predictability is at the nexus of the breach of the intentionality, causality or the consequence conditions. This is, thus, a key element to address when considering moral responsibility for AI. The next sub-section delves into this topic.

2.1 Predictability and AI Systems—What a System Will Do

Predictability of AI systems indicates the degree to which one can answer the question: *what will an AI system do?* This is not a new issue. Unpredictable systems are common in mathematics and physics, and limits on the ability to predict the outcomes of artificial systems have been proven formally since the 1950s (Moore, 1990; Musiolik & Cheok, 2021; Rice, 1956). Quite significantly, Wiener and Samuel debated over the predictability of AI systems in a famous exchange in 1960 (Samuel, 1960; Wiener, 1960). Wiener described the problems with the lack of predictability as a consequence of the learning abilities of these systems. As he noted, "as machines learn they may develop unforeseen strategies at rates that baffle their programmer" (Wiener, 1960, 1355). In baffling the programmer and the user, unpredictability hinders control, previsions about system behaviour, risks/benefit analyses, and undermines decisions made on those analyses (Taddeo & Floridi, 2018). It is important to note that the predictability problem refers both to correct and incorrect outcomes, as in both cases the issue is not whether the outcomes follow logically from the working of an AI system, but whether it is possible to foresee them at the time of deployment (Taddeo et al. Forthcoming). Developments in AI research have proved Wiener correct. Consider, for example, reward hacking—when a machine learning algorithm optimises an objective function, without achieving the intended outcome—which is reported in current literature as one of the reasons that make an AI system unpredictable. As Hadfield-Menell et al., (2020, 1) put it:

> Autonomous agents optimize the reward function we give them. […] When designing the reward, we might think of some specific training scenarios, and make sure that the reward will lead to the right behavior in *those* scenarios.

> Inevitably, agents encounter *new* scenarios (e.g., new types of terrain) where optimizing that same reward may lead to undesired behavior.

Currently, predictability of AI systems is a key focus of the debate on interpretable AI (Rudin, 2019*)*, which addresses problems related to (the lack of) predictability of AI systems in conjunction with the related, but distinct, issues of (lack of) transparency and explainability of these systems. There are different ways in which the literature on interpretable AI consider predictability, for example some focus on the technical features of a system (Boulanin et al., 2020; DIB, 2020b; International Committee of the Red Cross, 2019b), while others consider predictability a function of the systems and its context of deployment (Docherty, 2020; International Committee of the Red Cross, 2019a, 2019b).

When focusing on the technical aspects of an AI system, predictability is assessed with respect to the degree of consistency of its past and current behaviour with its future ones; by considering how often and for how long the outputs of a system are correct; and whether the system can scale up to elaborate data that diverge from training and test data (Boulanin et al., 2020; Collopy et al., 2020; DIB, 2020b).[2] Predictability varies with the type of AI systems—some, for example decision trees, are more predictable than others, like neural networks.

Predictability also depends on the robustness of the system. AI systems, especially those relying on online models, have low levels of robustness (Taddeo, 2017; Taddeo et al., 2019): they may show unpredicted, and unwanted, behaviour as a result of elaborating poisonous data presenting minor perturbations. For instance, in the case of image classification, perturbations at pixel-level, imperceptible to humans, can lead the system to misclassify an object with high-level confidence (Szegedy et al, 2014; Uesato et al., 2018). This implies that the number of possible perturbations that may alter the behaviour of an AI system is exorbitantly large and that predicting (and testing for) all possible perturbations to foresee unintended behaviour of a system is an intractable task.

Predictability also refers to the degree to which the actions of a system can be anticipated once it is deployed in a specific context. In this sense:

> all autonomous systems exhibit a degree of *inherent operational unpredictability*, even if they do not fail or the outcomes of their individual action can be reasonably anticipated, (Holland Michel, 2020, 5)(emphasis added).

This is because, once deployed, systems may face a context with unforeseen characteristics (IEEE, 2017). On this point, Holland Michel (2020) offers this example:

> a fully autonomous drone that maps the interior of a network of tunnels. Even if the drone exhibits a high degree of technical predictability and exceptional reliability, those deploying the drone cannot possibly anticipate exactly what it

[2] It is important to note that predictability is not reliability (the degree of failures of a system) nor is it robustness (the capacity of a systems to behave as expected even when it is fed with erroneous data) (Heaven 2019; Taddeo, McCutcheon, and Floridi 2019).

will encounter inside the tunnels, and therefore they will not know in advance what exact actions the drone will take.

The same point has also been stressed in an ICRC report (International Committee of the Red Cross, 2019a) with a specific focus on AWS:

> autonomous weapon systems are unpredictable in a broad sense, because they are triggered by their environment at a time and place unknown to the user who activates them. Moreover, developments in the complexity of software control systems – especially those employing AI and machine learning – may add unpredictability in a narrow sense of the process by which the system functions, (p. 11).

In this case, predictability is impacted by a large set of variables: the technical features of the system, the characteristics of the context of deployment, the level to which the operator understands the way in which the system works and, in the defence domain, the behaviour of the opponents. These variables may change and interact together in different ways making it problematic to predict all possible actions that an AI system may perform and their effects. As indicated in a UNIDIR report:

> Certain edge cases may only arise from a very specific and unforeseeable set of interacting circumstances. Testing a system against all such potential combinations of circumstances that might give rise to all of its possible failures could be extremely challenging (Holland Michel, 2020, 19).

Identifying all possible behaviours of AI systems in a given context and considering their effects on the context is not logically impossible, but it is unfeasible because the number of variables and their possible interactions is exorbitantly large, making this assessment intractable. This means that the deployment of AI systems implies a degree of uncertainty with respect to its possible effects. It is worth mentioning that uncertainty of outcomes may vary with the context of deployment. A system deployed in a context that is fully controllable is much more predictable than when deployed in a context where not all aspects are under control or known. Crucially, the lack of certainty of the outcomes introduces some risks which need to be pondered in the decision to use an AI system (see for example the principle of "justified use" proposed in Taddeo et al., 2021). The risks need to be assessed with respect to the perceived advantage of using AI, possible unwanted outcomes and mitigating measures put in place against these risks.

2.2 Collective and Faultless Distributed Moral Responsibility

Alternatives to the classic ethical approach to attribute human moral responsibilities for the actions of AI systems have been proposed in the relevant literature focusing either on collective (Corlett, 2001; List & Pettit, 2011) or distributed moral responsibility (Floridi, 2012, 2016).

Approaches focusing on collective responsibilities (Krishnan, 2009; List & Pettit, 2011) address the allocation of moral responsibilities for the actions taken by groups collectively, rather than their members individually. Consider, for example, List's and Pettit's analysis. It maintains that:

> […] a group agent is fit to be held responsible for doing something to the extent it satisfies these requirements:
> First requirement. The group agent faces a normatively significant choice, involving the possibility of doing something good or bad, right or wrong.
> Second requirement. The group agent has the understanding and access to evidence required for making normative judgments about the options.
> Third requirement. The group agent has the control required for choosing between the options. (List & Pettit, 2011, 158).

This analysis hinges on the premise that a group expresses its will and acts consequently. As the authors continue:

> To satisfy the second condition […], a group agent must be able to form judgments on propositions bearing on the relative value of the options it faces – otherwise it will lack normative understanding – and it must be able to access the evidence on related matters. (p. 158).

According to this approach, the group is seen as a homogenous entity—one may think for example of a group of workers striking, a group protesting in the street, the population of a nation—that acts in a coordinated manner having control over its actions and directing them intentionally. This is problematic when considering the design, development and deployment processes of AI systems, insofar as these are quite distributed and involve heterogeneous agents, who may not know of the overall action that the group is performing, let alone make normative judgments. By assuming an intentionality at group level, this approach disregards the impact that unintentional individual actions may have on the actions performed by the group. Thus, to rely on this approach would lead to unjust distribution of moral responsibility. As Corlett put it:

> collective (intentional) action is an action the subject of which is a *collective intentional agent*. A collective behavior is a doing or behavior that is the result of a collective, though not the result of its *intentions*. A collective action is caused by the beliefs and desires (wants) of the collective itself, *whether or not such beliefs and desires can be accounted for or explained in individualistic terms* […] I am concerned with whether or not it is justified to ascribe intentional action to conglomerates of a numerically larger sort such as (large) nations and (large) corporations. If such conglomerates are not intentional agents, then they are not proper subjects of moral responsibility attributions, (Corlett, 2001, 575–76) (emphasis added).

We agree with Corlett's analysis, ascribing intentionality to the group without being able to ascribe intentionality to all its agents undermines the idea of holding a group morally responsible, for those who did not share the intentions of the group would bear this responsibility without a justification.

The distributed morality approach focuses on the attribution of responsibility for morally evil/good actions that spur from the convergence of different, independent, morally neutral, intention-less factors. This has been defined as *distributed faultless moral responsibility* (Floridi, 2016). It refers to contexts in which, while it is possible to identify the causal chain of agents and actions that led to a morally evil/good outcome, it is not possible to attribute intent to achieve such an outcome to any of those agents individually, and therefore, all the agents are held morally responsible for that outcome insofar as they are part of the network that determined it. According to the distributed faultless moral responsibility approach, to attribute moral responsibility, what one needs to show is that:

> some evil has occurred in the system, and that the actions in question caused such evil, but it is not necessary to show exactly whether the agents/sources of such actions were careless, or whether they did not intend to cause them (Floridi, 2016, 8).

All the agents of the network are then held *maximally* responsible for the outcome of the network.

It is important to note that this approach allows for distributing moral responsibility among the human agents of a network but does not aim at distributing punishment or reward for the actions of a system. Its goal is to establish a feedback mechanism that incentivises all the agents in the network to improve its outcomes—if all the agents are morally responsible, they may become more cautious and careful, and this may reduce the risk of unwanted outcomes. This element is of particular relevance when considering AWS.

3 Moral Responsibility for AWS—the Collective Moral Responsibility Approach

When considering the specific case of AWS, attempts to overcome the responsibility gap have been made. Some build on the collective responsibility approach, others rely on the ideas of allocating responsibility along the chain of command or on distributed moral responsibility. In this section, we shall address these three attempts in turn, starting with those building on collective responsibility. Following the collective idea of moral responsibility, one may hold responsible this entire network of agents. For example, Taylor states:

> I maintain, the organisation as a whole might be viewed as having control over the outcome and thus be properly held morally responsible" (Taylor, 2020, 327).

This approach rests on the List's and Pettit's idea of collective responsibilities ascribed to groups of agents who share the intent to perform a given action. Indeed, it can be argued that those who work to design, develop and deploy AWS share the intent to develop a system that can deploy (lethal) force in a specific way and within certain constraints, and hence bear moral responsibility for the actions that the system is intended to perform. For the private sector, for example, this responsibility is akin to a liability that technology providers have with respect to the possible failure of their products. For example, Schulzke stresses that:

> To the extent that AWS' actions result from how their software or hardware is designed, responsibility for autonomous weapons should lie with the developers who create them. To the extent that their actions are enabled or constrained by civilian and military officials in their chain of command, those officials should share responsibility for the actions of autonomous weapons (Schulzke, 2013, 204).

This approach has been accepted in the past, when considering the development of other weapons (Weeramantry, 1985; Glerup & Horst, 2014; Miller, 2018; Khosrow-Pour D.B.A, 2021). However, it does not address the cases of AWS; insofar, it remains oblivious to the possibility that, once deployed, AWS may develop behaviour independently from the intentions of their designers/developers/users. The moral responsibility for this unintended behaviour is not akin to the one for systems' failure, as the latter can be discharged referring to negligence, i.e. the responsibility for some form of perturbation that the human agents could have and should have considered and prevented, but they did not. When considering unintended behaviour of AWS, we are focusing on behaviour that emerges as a correct consequence of the technical features of the systems (e.g. adapting systems) and/or their interactions with the surrounding environment but which could not be anticipated by the human agents, as mentioned in Sect. 2.1.

At the same time, the focus on the group leads to ascribing moral responsibilities, and blame/praise, to the organisations, rather than to individuals. As Taylor suggests:

> […] a number of distinct groups might be identified as potential loci of responsibility: the government, the military, and the developers of LAWS. Of these, I suggest that most progress in closing the responsibility gap can be made by exploring the possibility of ascribing responsibility to the organisations that design and develop LAWS (p. 327).

This approach offers a limited solution insofar as it is problematic to consider these groups *intentional agents* (the reader might recall Corlett's quote in Sect. 2.2). To overcome this issue, one may consider ascribing moral responsibility to individuals as representatives of groups or institutions (Champagne & Tonkens, 2015; Galliott, 2017). In this case, responsibility is attributed because of the role—and the obligations and duties attached to it—that an individual has and not necessarily because of their intentions or the factual connection between cause and effect (Hin-Yan Liu, 2016). Champagne and Tonkens, for instance, propose that persons of sufficiently high military or civilian standing hold responsibility for when the deployment of AWS goes wrong simply by virtue of their office (Champagne & Tonkens, 2015). By occupying high office, they argue, the occupant "willingly agrees" to the conditions of that office and, thereby, could in principle be made responsible for the use of AWS despite the latter's unpredictable nature.

In a sense, these proposals accord with the arrangements and values of a democratic system that the privileges of office are attended by certain responsibilities and liabilities for consequences not all of which are foreseeable or causally connected to the occupant of that office (Haddon, 2020). However, a closer analysis reveals that, at best, they ascribe moral responsibility *nominally*—responsible human agents are identified because of their role more than because of their intentions, decisions and actions, and also risks creating scapegoats and makes moral responsibility less meaningful. The reader may recall the quote from the Nuremberg trials, how it stressed the need to establish individual meaningful responsibility—"individuals who commit such crimes"—for wrongs occurred while waging warfare.

The next section analyses in more detail approaches focusing on the distribution of moral responsibility across the chain of command.

3.1 Moral Responsibility for AWS—Distributing Moral Responsibility Along the Chain of Command

Approaches that support the idea of distributing moral responsibility for the actions of AWS along the chain of command rest on two assumptions, according to which responsibility should be shared:

(i) proportionally to the decision-making power and the access to information characterising the different positions along the chain of command; and
(ii) according to the chain of command (within a military organisation) because of the authority that higher-ranking personnel have over the autonomy of lower-ranking personnel.

Assumption (i) resonates with the argument proposed by Walzer, according to which we have higher standards for commanders not just because of the dangerous instruments they have at their disposal but also because they have "access to all available information and also to the means of generating more information" (Walzer, 1977, 317). Assumption (ii) follows from (i) and is reflected in the "rules of engagement" (ROE) that defence organisations define prior to engaging in operations. ROE are established by cascading levels of hierarchy with each level imposing a greater degree of constraint—and specificity—upon its subordinate level(s). This means that, at each level, there is a (diminishing) degree of discretion and higher degree of specificity of possible decision/action, and it can be problematic to contradict the decisions made by superiors. Thus, personnel higher in command take responsibilities for actions performed by lower-rank personnel in executing their orders, because the latter lack the autonomy to do differently from what they are ordered. This view is shared by a good deal—albeit not all (McMahan, 2006)—of Just War theorising (Walzer, 1977). As Walzer's statement has it:

> we regard soldiers under orders as men whose acts are not entirely their own and whose liability for what they do is somehow diminished (p. 309).

This diminished autonomy, and thus diminished responsibility, is reflected in the doctrine of command responsibility, according to which superiors are held accountable for their subordinates by the principle of omission (e.g., a failure to prevent or intervene), when at least in principle they have sufficient control over their subordinate to prevent, or at least intervene to limit, immoral behaviour.

When considering AWS, assumptions (i) and (ii) are both problematic, insofar as they disregard the characteristics of AWS and the pragmatic and conceptual issues that may follow from their deployment. Let us consider assumption (i) first. It conflates the breadth of the decision-making power and of the information accessed with the level of granularity of the supporting information that is accessible. For example, in some circumstances, the risks and advantages of deploying AWS in a specific theatre of operation may be clearer to lower-ranking personnel, especially those familiar with the technology and the theatre, than to the higher ranking-personnel in command who may lack information specific to the technology or the context of deployment. And while information about risks and benefits of a specific operation may be conveyed upward in the chain of command, it is likely that in a context when AWS are deployed routinely and massively, the granularity of this information will decrease as the information is passed upward (Payne, 2021, 110–12). This may lead to personnel in higher command being held morally responsible for the actions of AWS, while having access to information of insufficient granularity for this responsibility to be attributed in a justified and fair way. It should be noted that this is a pragmatic problem. It is logically possible that military institutions may put in place adequate processes to overcome it. In this case, assumption (i) remains valid. But until a system is established to ensure that decision-makers have prompt access to sufficiently granular information about the benefits and risks of using AWS, the claim that the moral responsibility for the actions of AWS can be distributed with the level of command is unsound.

Assumption (ii) poses conceptual problems concerning control and autonomy. Let us consider them in turn, starting with control. The moral responsibilities of commanders for actions of their subordinates rest on three conditions:

> (1) the existence of a superior-subordinate relationship where the superior has effective control over the subordinate; (2) the requisite mental element, generally requiring that the superior knew or had reason to know (or should have known) of the subordinates' crimes; and (3) the failure to control, prevent or punish the commission of the offences (Neha Jain, 2016, 310).

Establishing control, whether "effective", "appropriate" or "meaningful", of AWS has proven to be problematic (Ekelhof, 2019). This may be in part due to the circumstances of deployment of AWS (consider for example the cases of humans on- or post-loop) or to the characteristics of the context of deployment, of the type of AWS or a combination of thereof. The reader may recall the issues of predictability discusses in Sect. 2.1, which entail that once deployed, AWS may perform unforeseen, and unintended, actions. This implies that officers deploying AWS may have no way to foresee unwanted outcomes and prevent them, making it hard to meet conditions (1) and (2). Condition (3) holds commanders responsible for the misconduct of their subordinates, but this is only insofar as they can control and prevent them from misbehaviour. In the case of AWS, their lack of predictability (whether technical or operational) makes the attribution of this moral responsibility unjustified.

3.2 Moral Responsibility for AWS—the Distributed Faultless Moral Responsibility Approach

Some of the limitations of the approaches described in the previous sections can be overcome when considering distributed faultless moral responsibility.

> According to this framework, commanders would be responsible for the actions of AWS to roughly the same extent as they are now, as they have similar powers to constrain the autonomy of AWS as they have over human soldiers. […] The exact apportionment of blame between commanders and developers can be determined only by the extent to which they contribute to an AWS' wrongful actions through their actions or inactions (Schulzke, 2013, 216) (p. 216).

This approach allows for allocating moral responsibility to all the individuals participating in (and determining) the design, development and deployment of AWS, but it would be problematic to ascribe moral blame or praise in a justified way following it. This is because of two reasons: lack of transparency and lack of intentionality.

The reverse engineering process necessary to identify the network of agents that shaped (causally) the behaviour of the AWS may be hindered by lack of transparency of the system itself or by the limited transparency of the information about the system and the decision-making process underpinning its use (Tsamados et al., 2021). The risk is concrete. It may be a consequence of the countless interactions among many agents that shape the actions of AWS, and which may be difficult to reconstruct with sufficient detail to understand the factors that determined the behaviour of the system. It can also follow the decision of a state not to share relevant information. In 2010, the UN Special Rapporteur on extrajudicial, summary or arbitrary executions stressed in a report on targeted killing that:

> […] [states may decide not to use] the procedural and other safeguards in place to ensure that killings are lawful and justified, and the accountability mechanisms that ensure wrongful killings are investigated, prosecuted and punished (Alston n.d., 10).

As Verdiesen, Santoni de Sio and Dignum stress in commenting on the report (Verdiesen et al., 2021):

> The reason for this accountability vacuum is that the international community cannot verify the legality of the killing, nor confirm the authenticity of the intelligence used in the targeting process or ensure that the unlawful targeted killing results in impunity (p. 145).

At the same time, as we saw in Sect. 2, reconstructing the causal chain of decision and actions that led to a specific behaviour of AWS is not tantamount to identifying intentions for that behaviour to materialise. Indeed, this approach aims at identifying *faultless* moral responsibility and does not aim to ascribe blame or praise. Thus, it sheds limited light on the responsibility gap of AWS; insofar, blame and praise are key elements to offer remedy for the moral evil that the use of these systems may cause or to reward for morally sound uses. The time has come to consider meaningful moral responsibility for the use of AWS.

4 Meaningful Moral Responsibility and The Moral Gambit

When considering AWS, moral responsibility refers to action leading to the destructive damage (whether lethal or not) that these systems may do. One of the conditions for the use of these systems to be morally acceptable is that it must be possible to ascribe responsibility for this damage in a justified way. This is when there is intentionality of the agents and a causal connection of their decisions/actions to the AWS outcomes. It is also important that moral responsibility is attributed fairly, such that the agents must have sufficient information and understanding of the context in which they operate to consider all possible alternatives before making any decision. Justified and fair moral responsibility is only attributed when all the four conditions specified in Sect. 2 are met. Also, the agents have to be able to accept this moral responsibility as an element of their actions and decisions, and have to be able to take the blame/praise that follows as an assessment of their moral character. *Meaningful* moral responsibility is ascribed meeting all these three criteria.

Discharging meaningful moral responsibility for the actions of AWS is a necessary, preliminary condition to their use,[3] because it is the kind of responsibility that shows a minimum due care (Strawson, 1962) for the receiver of the actions of AWS. In this sense, Sparrow is correct, when he says that:

> the least we owe our enemies is allowing that their lives are of sufficient worth that someone should accept responsibility for their deaths (Sparrow, 2007, 67).

Meaningful moral responsibility enables backward-looking responsibility, as it fosters accountability. It may also enable forward-looking responsibility, insofar as the prospect of the blame and praise linked to a given decision/action would facilitate morally sound choices and careful conduct.

When considering AWS, meaningful moral responsibility is limited by the unpredictable nature of these systems. It would be unjustified and unfair to ask human agents to take meaningful moral responsibility for (all possible) unpredictable actions that AWS may perform. This is because these actions are not intended by the human agent (moral responsibility would not be justified) and no information would allow the agent to foresee the totality of possible actions that an AWS may perform once deployed, to identify and prevent those unwanted (moral responsibility

[3] Please note that we are not arguing that moral responsibility is necessary to ensure that war activities abide to Just War Theory or that it is necessary to ensure the respect of International Humanitarian Law. We argue that attributing moral responsibility is necessary to ensure that war waging is *morally sound*.

would not be fair). All one may ask is for designers, developers and deployers to take meaningful moral responsibility for the intended actions, *while* being aware of the risk that unpredicted outcomes may occur and *accepting* moral responsibility also for the unpredictable effects that may follow the decision to deploy AWS. Let us specify this aspect. In accepting this responsibility, the human agents make a *moral gambit*: they design/develop/use an AWS, being fully aware of the risks that it may perform some unpredicted actions. To limit these risks (and optimise the chances for a successful gambit) they, and the relevant defence institutions, must act at their best to establish all possible measures to constrain the moral evil (and harness the moral good) that unpredicted behaviour may cause. The human agents remain aware that independently of all these efforts, it will not be feasible to predict all possible actions of an AWS and their effects on the context of deployment.[4] Nonetheless, if they decide to proceed with the design/development/use of these systems, then they make a moral gambit and accept to be morally responsible for the unforeseen AWS outcomes and their effects.

When considering AWS, all the human agents intentionally participating in the design, development and use of these systems take this moral gambit. Personnel who decide on the deployment of these systems take (or not) the gambit, only insofar as they have a choice as to use/not use AWS. Hence, the responsibility gap for the actions of AWS may be closed so long as we can identify intentionality, causality with respect to the behaviour of AWS, full awareness about unpredictability of these systems and willingness to take the moral gambit. As we will see in the next section, for this to be possible, an infrastructure for accessing relevant information, to ensure traceability of processes, and non-lethal outcomes needs to be established. Before moving to the next section, three clarifications are necessary to clarify the boundaries of the gambit.

The first clarification addresses the scope of the gambit. The moral gambit does not concern the decision to participate in an operation whose outcomes are relatively unpredictable. Rather, it is about the *voluntary acceptance* of moral responsibility for the range of possible outcomes that may follow the use of AWS, whether foreseen or not. In this sense, the moral gambit can be neither mandated, it must be taken in a voluntary way and the responsibility that comes with it is *accepted*, not attributed. In this sense, the moral gambit is about accepting ex ante responsibility for whatever may happen in that specific operation, hoping that unintended and unwanted or unforeseen outcomes never occur, but also accepting being held responsible if they do.

This takes us to the second point to clarify: the approach underpinning the gambit. In a context where intentionality and effects of an action are separated, moral responsibility risks either to be voided of any meaning—e.g. when it is ascribed nominally—or to be non-ascribable, as per the responsibility gap. The moral gambit overcomes this dilemma by shifting the focus from *attributing* moral responsibly to *accepting* it on a voluntarily basis. This approach places a

[4] Although it is worth noting that the extent of, at least technical, unpredictability of these actions and effects may be constrained within predetermined boundaries such as payload or range.

heavier burden on the human agents to accept the gambit and the moral responsibility that comes with it. This is why those who accept to take the gambit need to be fully informed and aware of risks and consequences of their choices, and institutions have a fundamental duty to support these individuals (more on this in the next section). But this heavier weight offers a way to overcome the responsibility gap, while ensuring that the moral responsibility remains meaningful and fair.

The third clarification concerns the permissibility of the moral gambit. In the context of national defence, when considering non-lethal AWS, the moral gambit may be acceptable. This is not the case when considering LAWS. In this case, the moral gambit would be a gambit taken on the lives of others. This is why, on the basis of our analysis, we argue that even if a human agent were willing to accept this gambit, it should not be permitted and would remain morally unacceptable.

The impermissibility of taking the moral gambit on the lives of others breaks down in two ways depending on the recipient of acts of war. Pertaining to non-combatants, the moral gambit is straightforwardly ruled out by the principle of distinction. Distinction provides non-combatants with immunity from attack in times of war at all times. An unpredictable system, with regard to its ability to identify, select and engage legitimate human targets, does not respect the principle of distinction (Blanchard & Taddeo, 2022a).

The impermissibility of the moral gambit on the lives of combatants is more complex. It is worth stressing that, in this case, the problem is with the mode of killing and not with the killing per se (Blanchard & Taddeo, 2022b). One could object (Walzer, 1977, 42; Tamar Meisels, 2018, 11–29) that combatants forgo their right not to be killed, and therefore, it makes no difference if they are killed by humans or by LAWS. Whether combatants waive their right to life is a controversial point (Bazargan, 2014; Kamm, 2004; McMahan, 2011), but let us accept it for the sake of argument.

The objection here is misplaced, because the issue here is not *whom* is liable of being killed, but *how* it is acceptable to kill those liable to be killed. Combatants may waive their right to life, but they do so under expectation that an attack on their life will comply with the principle of military necessity and respect the principle of the moral equality of combatants. Now, it is conceivable, though implausible, that LAWS may be used against human combatants in a way which complies with military necessity. It is conceivable because the Just War doctrine of “supreme emergency” justifies the use of any means if the stakes are high enough to warrant them—i.e. survival (Walzer, 1977, 251–68). It is implausible both because such circumstances seldom exist and because the operational advantages imputed to LAWS—such as decision-making speed—far exceed the capabilities of humans, thereby lowering the threshold for the necessity of using them against human combatants.

As for moral equality of combatants, LAWS breach this principle. Under the moral equality of combatants, combatants are said to enter into a “martial contract” entailing “mutual belligerency rules” whereby the waiving of the right not to be killed is attended by a system of norms. As Skerker et al. write:

> Service members can be modelled as ceding claim-rights against being targeted with lethal violence by enemy combatants according to the terms of military norms optimizing a balance between maximising the one military's interests whilst minimizing suffering to their enemy (Skerker et al., 2020, 202).

One of those norms is the expectation that combatants will not be targeted in a way that is either wanton or cavalier. This implies that unintended killing (whether of combatants or not) is at the very least morally problematic, because even if combatants waive their right not to be attacked, they do so under the assumption that their lives will be taken intentionally, not as a result of an unintended behaviour or a lost moral gambit. Personnel choosing to deploy a LAWS would *take a gambit* that it will act according to the intended use, while being aware that there is a chance that a LAWS may identify, select and engage an unintended target. This gambit, we argue, is incompatible with the expectations imposed by the principles of necessity and moral equality of combatants. This is why, insofar as LAWS are unpredictable with regard to the identification, selection and engagement of human targets, it is not possible to ascribe meaningful moral responsibilities for their actions.

5 Discharging Meaningful Moral Responsibility for the Actions of Non-Lethal AWS

Much of the debate on AWS centres on LAWS, as lethal uses of AWS pose high-level risks to life and thus severe ethical problems. As such, non-lethal AWS have attracted less attention. We believe that this gap is problematic. While the ethical risks posed by non-lethal AWS may be less "dramatic" than LAWS, these remain nevertheless serious ethical risks. Non-lethal AWS can cause significant damage, including bodily harm, disproportionate destruction to property, infringements of liberty and breaches of the principle of distinction. These ethical risks arise for example when considering the use of non-lethal AWS for national security purposes "when the use of graduated force is required and deadly force is the exception" (Heyns, 2016a, 5). Thus, attributing moral responsibility for the use of non-lethal AWS is crucial. The proposed moral gambit offers a new and much-needed solution in this respect.

The moral gambit cannot be imposed upon a human agent, nor can it "come with the role" that one may acquire. For it to be acceptable, the agent must make it willingly. When considering the use of *non-lethal AWS*, there are a number of procedures that can be established to support the decision to take, or not to take, the moral gambit. In what follows, we offer eight recommendations that military organisations should follow to support their members in this sense. It should be noted that albeit addressing military organisations specifically, the recommendations also bear on technology providers. Also, the permissibility of the moral gambit for non-lethal LAWS hinges on the calculation of the risks for unintended lethal outcomes and the high-level assurance that, given design specifications and deployment conditions, the risk of this outcome has been reduced to the lowest possible level.[5] The recommendations are listed in logical order.

1. Procure AWS whose underlying AI systems are *interpretable* and not just explainable. Lack of predictability of AI systems is also a function of their lack of transparency. Explanations of a black box model may offer imprecise representation of the original model (Rudin, 2019); for this reason, explainability offers limited solutions to the problems posed by lack of predictability. Better outcomes may be reached by procuring interpretable models, i.e. a model, that is:

   > constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity or physical constraints that come from domain knowledge, (Rudin, 2019, 206).

2. Assess predictability. Providers and defence institutions should assess which technical and operational features of AWS encroach upon the predictability of their ability to identify, select and engage targets and work to improve them to limit unpredictable outcome.
3. High level of knowledge and understanding for the decision-makers. Those deciding about the use of non-lethal AWS should have a high level of technical understanding of these systems and of the theatre of operation, so as to be able to identify their strategic and tactical potential, how to deploy them best, as well as their possible points of failure. High level of technical understanding underpins the choice to use non-lethal AWS insofar as it enables the decision-makers to consider the risks of unpredictable behaviour properly and of the implication this has for their moral gambit.
4. Traceability of processes. Information about the technical specification of a non-lethal weapons system, its cycle of development and mode of procurement should be transparent to the decision-makers. In the same way, any relevant information about the system that may advance the understanding that the decision-maker should be relayed to the personnel promptly and accurately.
5. Justification of uses. The decision as to use or not use non-lethal AWS should always follow a risk/benefit analysis and be justified according to the principle of necessity. A further consideration should also be made about the safety of military personnel deployed in the specific theatre.
6. Ensuring non-lethal effects. Measures must be put in place to minimise the risks of lethal outcomes from the use of non-lethal AWS. These may not differ too much from similar measures taken when using conventional weapons and could include a range of approaches such as necessity and proportionality calculations, assessment of the theatre of operation and user interface design features like a remote switch button.

[5] Further work should be developed on acceptable level for this risk.

7. Redressing and remedy. A process to identify mistakes and unwanted outcomes, to assess their impact and costs and to define redressing remedy measures should be established. Redressing and remedy measures will not override the moral praise or blame attributed to human agents, but should be used to discharge the accountability that the defence institutions has with respect to the decisions made by its personnel.
8. Auditing. Ethics-based auditing of both the non-lethal AWS and of the processes for their acquisition and deployment should be established (Mökander & Floridi, 2021), with the aim of facilitating accountability as well as to identify possible points of failure and address them promptly, so to improve the decision-making and the redressing processes.

6 Conclusions

In this article, we focused on the attribution of moral responsibility for the actions of AWS. We argued that for the use of these systems to be ethically acceptable, it is crucial to attribute *meaningful* moral responsibility to human agents. The attribution of this kind of moral responsibility rests on strong requirements, which are justified because of the nature of the damage that AWS may cause (whether lethal or not). We argue that these requirements cannot be met when considering the case of LAWS (as defined by Taddeo and Blanchard, Forthcoming) and, thus, that the deployment of these systems is morally unacceptable. We also stress that meaningful moral responsibility can be accepted by means of the moral gambit for the actions of non-lethal AWS, provided that defence institutions establish necessary processes to support human agents willing to take the gambit understanding of the functioning of AWS and of the benefits and risks linked to their uses, and the consequences of the moral gambit. We hope that the eight recommendations provided in this article will help technology providers and defence organisations to this end.

Funding Mariarosaria Taddeo and Alexander Blanchard have been funded by the Dstl Ethics Fellowship held at the Alan Turing Institute. The research underpinning this work was funded by the UK Defence Chief Scientific Advisor's Science and Technology Portfolio, through the Dstl Autonomy Programme, grant number R-DST-TFS/D026.

Declarations

Ethical Approval and Consent No human or animal subject was involved in the research related to this paper; hence, ethical approval, consent to participate and consent to publish do not apply to this article.

Competing Interests The authors declare no competing interests.

Disclaimer This paper is an overview of UK Ministry of Defence (MOD) sponsored research and is released for informational purposes only. The contents of this paper should not be interpreted as representing the views of the UK MOD, nor should it be assumed that they reflect any current or future UK MOD policy. The information contained in this paper cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

References

Alston, P. n.d. 'Report of the special rapporteur on extrajudicial, summary or arbitrary executions, Philip Alston: Addendum - Study on Targeted Killings (A/HRC/14/24/Add.6) - Russian Federation'. ReliefWeb. Accessed 5 June 2021. https://reliefweb.int/report/russian-federation/report-special-rapporteur-extrajudicial-summary-or-arbitrary-executions.

Bazargan, S. (2014). Killing minimally responsible threats. *Ethics, 125*(1), 114–136. https://doi.org/10.1086/677023

Bentham, J. (1789). *An introduction to the principles of morals and legislation.* Garden City: Doubleday.

Blanchard, A., & Taddeo, M. (2022). Predictability, distinction & due care in the use of lethal autonomous weapons systems. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4099394

Blanchard, A., & Taddeo, M. (2022). Autonomous weapon systems and jus ad bellum. *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01425-y

Boulanin, V., Carlsson M. P., Goussac, N., & Davidson, D. (2020). 'Limits on autonomy in weapon systems: Identifying practical elements of human control'. Stockholm International Peace Research Institute and the International Committee of the Red Cross. https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0.

Branscombe, N. R., Owen, S., Garstka, T. A., & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome?1. *Journal of Applied Social Psychology, 26*(12), 1042–1067. https://doi.org/10.1111/j.1559-1816.1996.tb01124.x

Champagne, M., & Tonkens, R. (2015). Bridging the responsibility gap in automated warfare. *Philosophy and Technology, 28*(1), 125–137.

Coleman, S. (2015). Possible ethical problems with military use of non-lethal weapons international regulation of emerging military technologies. *Case Western Reserve Journal of International Law, 47*(1), 185–200.

Collopy, P., Sitterle, V., & Petrillo, J. (2020). Validation testing of autonomous learning systems. *Insight, 23*(1), 48–51. https://doi.org/10.1002/inst.12285

Corlett, J. A. (2001). Collective moral responsibility. *Journal of Social Philosophy, 32*(4), 573–584. https://doi.org/10.1111/0047-2786.00115

Davison, N. (2009). *'Non-lethal' weapons*. Palgrave Macmillan.

DIB. (2020a). 'AI principles: Recommendations on the ethical use of artificial intelligence by the department of defense'. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

DIB. (2020b). 'AI principles: recommendations on the ethical use of artificial intelligence by the department of defense - Supporting document'. Defense Innovation Board [DIB]. https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF.

Docherty, B. (2020). 'The need for and elements of a new treaty on fully autonomous weapons'. *Human Rights Watch*, 1 June 2020. https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348. https://doi.org/10.1111/1758-5899.12665

Enemark, C. (2008). "Non-lethal" weapons and the occupation of Iraq: Technology, ethics and law. *Cambridge Review of International Affairs, 21*(2), 199–215.

Enemark, C. (2008). "Non-lethal" weapons and the occupation of Iraq: Technology, ethics and law. *Cambridge Review of International Affairs, 21*(2), 18.

Fischer, J M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. First paperback ed. Cambridge Studies in Philosophy and Law. Cambridge: Cambridge University Press.

Floridi, L. (2012). Distributed morality in an information society. *Science and Engineering Ethics, 19*(3), 727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L., & Taddeo, M. (2018). Romans would have denied robots legal personhood. *Nature, 557*(7705), 309–309. https://doi.org/10.1038/d41586-018-05154-5

Galliott, J. (2017). *Military robots: Mapping the moral landscape*. http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781317096009.

Glerup, C., & Horst, M. (2014). Mapping "social responsibility" in science. *Journal of Responsible Innovation, 1*(1), 31–50. https://doi.org/10.1080/23299460.2014.882077

Haddon, C. (2020). 'Ministerial accountability'. The Institute for Government. 16 September 2020. https://www.instituteforgovernment.org.uk/explainers/ministerial-accountability.

Hadfield-Menell, Dylan, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. 2020. 'Inverse reward design'. ArXiv:1711.02827[Cs] , October.http://arxiv.org/abs/1711.02827

Heaven, D. (2019). Why Deep-Learning AIs Are so Easy to Fool. *Nature, 574*(7777), 163–166. https://doi.org/10.1038/d41586-019-03013-5

Heyns, C. (2016a). 'Autonomous weapons systems: living a dignified life and dying a dignified death'. In *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, and Claus Kreß. Cambridge: Cambridge University Press.

Heyns, C. (2016). Human rights and the use of autonomous weapons systems (AWS) during domestic law enforcement. *Human Rights Quarterly, 38*(2), 350–378. https://doi.org/10.1353/hrq.2016.0034

Hin-Yan, L. (2016). 'Refining responsibility: Differentiating two types of responsibility issues raised by autonomous weapons systems'. In *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, and Claus Kreß. Cambridge: Cambridge University Press.

Holland Michel, A. (2020). 'The black box, unlocked: Predictability and understandability in military AI'. United Nations Institute for Disarmament Research. https://doi.org/10.37559/SecTec/20/AI1.

IEEE. (2017). 'Reframing autonomous weapons systems'. IEEE. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_reframing_autonomous_weapons_v2.pdf.

International Committee of the Red Cross. (2019a). 'Artificial intelligence and machine learning in armed conflict: A human-centred approach | International Committee of the Red Cross'. https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach.

International Committee of the Red Cross, ICR. (2019b). 'Autonomy, artificial intelligence and robotics: Technical aspects of human control'. https://www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

International Military Tribunal (Nuremberg). (1947). Judgment and sentences, October 1, 1946. *American Journal of International Law, 41*, 172–306.

Kamm, F. M. (2004). Failures of just war theory: Terror, harm, and justice. *Ethics, 114*(4), 650–692. https://doi.org/10.1086/383441

Kant, I., & Borken, T. (2019). *Grundlegung zur Metaphysik der Sitten (Großdruck)*. https://nbn-resolving.org/urn:nbn:de:101:1-2019040502040341963072.

Kaurin, P. M. S. (2010). With fear and trembling: An ethical framework for non-lethal weapons. *Journal of Military Ethics, 9*(1), 100–114. https://doi.org/10.1080/15027570903523057

Kaurin, P. M. S. (2015). And next please - The future of the NLW debate international regulation of emerging military technologies. *Case Western Reserve Journal of International Law, 47*(1), 217–228.

Kelly, E. I. (2012). 'What is an excuse?' In *Blame*, edited by D. Justin Coates and Neal A. Tognazzini, 244–62. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199860821.003.0013.

Khosrow-Pour D.B.A., Mehdi, ed. (2021). *Encyclopedia of Information Science and Technology, Fifth Edition:* Advances in information quality and management. IGI Global. https://doi.org/10.4018/978-1-7998-3479-3.

Khoury, A. C. (2018). The objects of moral responsibility. *Philosophical Studies, 175*(6), 1357–1381. https://doi.org/10.1007/s11098-017-0914-5

Krishnan, A. (2009). *Killer robots: Legality and ethicality of autonomous weapons*. Ashgate.

Lebreton, G. (2021). 'Report of the committee on legal affairs to the EUropean Parliament'.

Levy, N. (2008). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, and Psychology, 14*(2), 129–138. https://doi.org/10.1353/ppp.0.0003

List, C., & Pettit, P. (2011). Group Agency. *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780199591565.001.0001

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McMahan, J. (2006). On the moral equality of combatants. *Journal of Political Philosophy, 14*(4), 377–393.

McMahan, J. (2011). Who is morally liable to be killed in war? *Analysis, 71*(3), 544–559.

Miller, S. (2018). *Dual Use science and technology, ethics and weapons of mass destruction*. New York, NY: Springer Berlin Heidelberg.

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09557-8

Moore, C. (1990). Unpredictability and undecidability in dynamical systems. *Physical Review Letters, 64*(20), 2354–2357. https://doi.org/10.1103/PhysRevLett.64.2354

Musiolik, TH, & AD Cheok, eds. (2021). *Analyzing future applications of AI, sensors, and robotics in society:* Advances in Computational intelligence and robotics. IGI Global. https://doi.org/10.4018/978-1-7998-3499-1

Neha, J. (2016). 'Autonomous weapons systems: New frameworks for individual responsibility'. In *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by Nehal Bhuta, Susanne Beck, Robin Geiß, Hin-Yan Liu, and Claus Kreß. Cambridge: Cambridge University Press.

Nelkin, D., K. (2011). *Making sense of freedom and responsibility*. Oxford ; New York: Oxford University Press.

Payne, K. (2021). *I, warbot: The dawn of artificially intelligent conflict*. Hurst & Company.

Rice, H. G. (1956). On completely recursively enumerable classes and their key arrays. *Journal of Symbolic Logic, 21*(3), 304–308. https://doi.org/10.2307/2269105

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samuel, A. L. (1960). Some moral and technical consequences of automation–A refutation. *Science, 132*(3429), 741–742. https://doi.org/10.1126/science.132.3429.741

Sartorio, C. (2007). Causation and responsibility. *Philosophy Compass, 2*(5), 749–765. https://doi.org/10.1111/j.1747-9991.2007.00097.x

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy and Technology, 26*(2), 203–219. https://doi.org/10.1007/s13347-012-0089-0

Shoemaker, D. (2017). *Oxford Studies in Agency and Responsibility 4 4*.

Skerker, M., Purves, D., & Jenkins, R. (2020). Autonomous weapons systems and the moral equality of combatants. *Ethics and Information Technology, 22*(3), 197–209. https://doi.org/10.1007/s10676-020-09528-0

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Strawson, P. (1962). Freedom and resentment. *In Proceedings of the British Academy, 48*(1962), 1–25.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). 'Intriguing properties of neural networks'. ArXiv:1312.6199[Cs] , February. http://arxiv.org/abs/1312.6199

Taddeo, M. (2017). Trusting digital technologies correctly. *Minds and Machines, 27*(4), 565–568. https://doi.org/10.1007/s11023-017-9450-5

Taddeo, M., & Blanchard, A. (2021). 'A comparative analysis of the definitions of autonomous weapons systems'. Academic report. Geneva, Switzeland: UN GGE CCW.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–52. https://doi.org/10.1126/science.aat5991

Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence, 1*(12), 557–560. https://doi.org/10.1038/s42256-019-0109-1

Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2021). Ethical principles for artificial intelligence in national defence. *Philosophy and Technology*. https://doi.org/10.1007/s13347-021-00482-3

Taddeo, M., Ziosi, M., Tsamados, A., Kurapati, S., & Gilli, L. (n.d.) Forthcoming. 'Artificial intelligence for national security: The Predictability Problem'. Alan Turing Institute.

Tamar, M. (2018). *Contemporary just war: Theory and practice*. Routledge.

Taylor, I. (2020). Who is responsible for killer robots? Autonomous weapons, group agency, and the military-industrial complex. *Journal of Applied Philosophy n/a (n/a)*. https://doi.org/10.1111/japp.12469

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: Key problems and solutions. *AI and Society*. https://doi.org/10.1007/s00146-021-01154-8

Uesato, J., O'Donoghue, B., van den Oord, A., & Kohli, P. (2018). 'Adversarial risk and the dangers of evaluating against weak attacks'. ArXiv:1802.05666[Cs, Stat], February. http://arxiv.org/abs/1802.05666

UN GGE CCW. (2019). 'Group of governmental experts on emerging technologies in the area of lethal autonomous weapons system, (2019). Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. Geneva: The United Nations Office at Geneva.' Geneva: The United Nations Office at Geneva.

Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: A framework for comprehensive human oversight. *Minds and Machines, 31*(1), 137–163. https://doi.org/10.1007/s11023-020-09532-9

Wallace, R. Jay. 1998. *Responsibility and the moral sentiments*. 2. print. Cambridge, Mass.: Harvard Univ. Press.

Walzer, M. (1977). *Just and unjust wars: A moral argument with historical illustrations*. Basic Books.

Watson, G. (1975). Free agency. *The Journal of Philosophy, 72*(8), 205. https://doi.org/10.2307/2024703

Weeramantry, C. G. (1985). Nuclar weaponary and scientific responsibility. *Journal of the Indian Law Institute, 27*(3), 351–386.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.