Predictability, Distinction & Due Care in the use of Lethal Autonomous Weapon Systems

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

[1] The Alan Turing Institute, London, UK

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

Abstract

In this article we address the possibility of using Lethal Autonomous Weapon Systems (LAWS) in compliance with the *jus in bello* principle of distinction. This principle requires that parties to an armed conflict distinguish between military objects and civilian objects, directing their actions only towards the former. We argue that systems characterised by unpredictability cannot be used in a way which respects the principle of distinction and that this is particularly problematic when these systems target human agents. We find that the use of LAWS unacceptably weights the risks of conflict towards civilians, in breach of the obligation of due care, which states that combatants are obliged to accept greater risks to themselves to ensure that they hit only the right target, and that there is not the military necessity warranting an infringement of this obligation. We develop our analysis through a critique of a recent International Committee of the Red Cross position on LAWS.

Keywords: Artificial Intelligence, Jus in Bello Distinction, Due Care, Predictability, Lethal Autonomous Weapons Systems, Military Ethics, Just War Theory.

Funding information: Alexander Blanchard and Mariarosaria Taddeo's work on this article has 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. 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

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

contained in this paper cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

Acknowledgements: The authors wish to thank Christopher J. Finlay for his helpful comments on the manuscript.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Predictability, Distinction & Due Care in the use of Lethal Autonomous Weapon Systems

1. Introduction

There has been ample scholarly and international debate about the ethical implications of the use of Lethal Autonomous Weapon Systems (LAWS). The question of whether LAWS should or should not be subject to a ban hinges in large part on their ability to respect the principle of distinction.[1] The principle of distinction requires that parties to an armed conflict distinguish between military targets and civilian targets, directing their actions only towards the former.

Some commentators have argued that LAWS are able to respect the principle of distinction in a way which surpasses human combatants. LAWS are said to do so for three reasons. First, not possessing a sense of self-preservation, LAWS can act in a more conservative manner when there is a low certainty of target identification. A principle of 'first do no harm', as Arkin writes, would allow LAWS to "truly assume risk on behalf of the non-combatant" (Arkin 2018, 3). Second, since LAWS are free of 'fear and hysteria', they are purported to cut through the fog of war, reducing instances of atrocities and war crimes (Marchant et al. 2011, 280; Grut 2013, 11). Third, LAWS are not prone to problems of 'scenario fulfilment' whereby, in stressful situations, humans fit or distort information into familiar patterns (Arkin 2009). As such, able to assimilate contradictory data, LAWS are expected to reduce instances of target misidentification. It is because of this promise of comparatively greater protection for civilians that a United States Congress commission declared a "moral imperative" to develop AI weapon systems (Jeffrey Dastin and Paresh Dave 2021). More broadly, one may argue that as LAWS are endowed with AI systems (particularly AI systems implanting online learning models, more on this presently), they can rely on autonomy and learning capabilities that improve performances, lead to more effective outcomes as well as overall support for a military force in gaining (or maintaining) an advantage over its adversaries, while also protecting its own personnel. The same adaptive capabilities, however, also pose serious ethical problems concerning the attribution of moral responsibilities for the actions of LAWS (Taddeo and Blanchard Forthcoming), breaches of the Just War Theory principle of *in bello* necessity (Blanchard and Taddeo Forthcoming) and, as we argue in

[1] Often also referred to as the principle of discrimination.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

this article, a breach of the principle of distinction. These ethical problems stem from the lack of (limited) predictability of the outcomes generated by LAWS.[2]

Lack of predictability, or limited predictability, characterises systems with learning and adaptive capabilities. As Roff and Moyes write, predictability is a primary metric by which "humans can measure whether their creations are continuing to function as intended" (Roff and Moyes 2016, 2). An unpredictable system poses challenges for anticipating its effects, for ensuring its explainability, reliability, and trustworthiness, and poses broader challenges for ascribing responsibility should the system cause unintended harm (Taddeo et al. 2021).

The ethical problems posed by unpredictable systems are exacerbated in military uses where unique challenges, related to the use of force, compound problems already present in civilian use of these systems. Dynamic environments, greater risk profiles, adversarial behaviour, and system complexity, make predictability more critical and yet harder to achieve (Horowitz 2019). Additionally, the unpredictability of LAWS poses questions as to whether they can be used in a way which adheres with International Humanitarian Law (IHL) and its principles of necessity, proportionality, and distinction. For example, unpredictability may hinder proportionality calculi through diminishing the foreseeability of certain costs in employing LAWS. Likewise, it will be difficult to use LAWS in a way which respects the principle of distinction between combatants and non-combatants if the system cannot be relied upon to behave in the way intended by its deployer – i.e. to achieve the intended purpose of use.

Indeed, because of the challenges in respecting the principle of distinction there have been calls either for a moratorium or an outright ban on the use of LAWS altogether. For instance, the International Committee of the Red Cross (ICRC) recommends that,

> "Unpredictable autonomous weapon systems should be expressly ruled out, notably because of their indiscriminate effects. This would best be achieved with a prohibition on autonomous weapon systems that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained" (ICRC 2021, 2).

In this article we argue that unpredictable LAWS – i.e. a LAWS for which the outputs of each and all functions that enable the system to identify, select and engage a target are not predictable – cannot be used in a way that respects the principle of distinction, and that therefore these systems cannot be

[2] See the chairperson's summary for the April 2021 meeting of the UN Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems of the Convention on Certain Conventional Weapons (UN CCW GGE 2021).

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

used justifiably against human agents. In making this argument, we agree with the sentiment behind the ICRC position, but we argue it has two limitations that should be addressed to make it more compelling.[3]

First, ICRC provides an inadequate explication of the principle of distinction to constitute a solid basis for recommending a ban on LAWS. In its narrowest interpretation the principle of distinction prohibits only the *intentional* harming of non-combatants. We argue that any prohibition of LAWS based on the principle of distinction must refer to the associated ethical obligation of due care to provide adequate grounds for capturing what is morally problematic about LAWS.

Second, the position of the ICRC is too permissive because it fails to account for the contextual awareness required by the operator of a weapon system for determining the 'proper target' prior to the application of force by that system. Thus, the ICRC recommends a ban at most on the *application of force* by LAWS, leaving unaddressed a whole range of targeting functions which remain permissible under the ICRC ban. We argue that this approach is problematic for being incompatible with the obligation of due care.

In the rest of this article, in section 2 we provide a characterisation of predictability as it relates to LAWS before recounting the position of the ICRC on LAWS. In section 3, we provide a full account of the principle of distinction as supplemented by the obligation of due care. In section 4, we address the problem of predictability for undertaking *in bello* assessments of liability to attack, before concluding the article in section 5.

Before beginning we must define what we mean by an autonomous weapon system (AWS):

> "…an artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal, or set of goals, within its dynamic operating environment and without the direct intervention of another agent and may also be endowed with some abilities for changing its own transition rules[4] without the intervention of another agent, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object

[3] Whilst the ICRC basis its recommendation on an analysis of "humanitarian, legal, ethical, technical and military implications of AWS," (ICRC 2021, 3) it refers to the principle of distinction in the context of International Humanitarian Law (IHL). Here our discussion of the principle of distinction is drawn from the ethical framework of Just War Theory. The novelty of the challenges posed by LAWS necessitates legal analyses be coupled with conceptual and ethical analyses to grasp most accurately the challenges presented. Indeed, there are significant overlaps between Just War Theory and IHL, and we hope that these ethical considerations might help strengthen the regulatory recommendations of the ICRC.

[4] Transition rules allow a Turing Machine, and more in general computational artefacts, to change its internal state when presented with a certain output following a series of if/then or logic doors as defined by its algorithm. An internal state is the internal configuration of a computational artefact, it encompasses the variables (and their weights) that affect the behaviour of the artefact.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

> or a human 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, 19).

This definition has a number of advantages. Firstly, it is both comparative and value-neutral. Being comparative, it draws on various state definitions and can be used to underlay international cooperation on the regulation of LAWS, particularly where state definitions diverge. Being value-neutral it provides nothing further to our analysis than the definition of the set of systems to which our analysis applies. Secondly, it permits the categorization of AWS according to purpose of use. Here, LAWS (lethal AWS) are considered a subset of AWS because they are used for the purpose of exerting lethal force against human beings, that is, for the purpose of either killing or inflicting permanent injury. Likewise, non-lethal AWS are AWS used for non-lethal purposes, that is "without causing death or permanent injury" (Davison 2009, 1). A non-lethal AWS might be used for anti-material purposes, such as the destruction of an incoming projectile, or for the disabling of equipment. Our argument applies only to LAWS as we define it. It is worth noting this definition of LAWS differs from many others insofar as it considers explicitly adapting capabilities as a key feature of LAWS. As we mentioned earlier, it is this characteristic which drives much of the discussion in this article and around the attribution of moral responsibility more generally for the actions of LAWS.

2. Characterising Predictability

Predictability is an important component for the ethical use of a system. If a system's behaviour is unpredictable, that hinders our capacity to trust and to rely on that system (Taddeo 2017; 2019; Taddeo, McCutcheon, and Floridi 2019). Predictability can refer to the characteristics of the system and the degree of certainty to which its behaviour can be predicted (technical predictability). It can also refer to the interaction of the system with its specific context of deployment and the degree of certainty with which the behaviour of the system can be anticipated in that specific context (operational predictability).

When considering the technical aspects of an AI system, predictability is assessed according to three criteria: whether there is consistency between past, present, and future behaviours; whether the frequency and duration of the outputs of the system are correct; whether the system can scale up data utilized beyond its testing stage of development (Boulanin et al. 2020; Collopy, Sitterle, and Petrillo 2020; DIB 2020). The predictability of LAWS in a technical sense will vary from system to

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

system. LAWS can entail a wide range of AI models with varying degrees of autonomy, assigned to specific tasks which, individually, may or may not be predictable. An AI model which employs decision trees, for example, will be more predictable than one which employs a neural network. Even between examples of the same type of model, there will be varying degrees of predictability. For instance, machine-learning models which have offline learning modes will be more predictable than models employing online reinforcement learning modes throughout the duration of deployment. This is because online reinforcement learning collapses the distinction between training, testing, and operational use of the system. As Mattias writes,

> "The system learns inside its final operating environment by exploring available action alternatives in a trial-and-error fashion, optimising its own parameters according to the results. Thus, the exploration phase is an integral part of the design of the working machine and cannot be separated from it" (Matthias 2004, 176)

Offline learning models are more predictable but also less innovative and may offer poorer performance than cutting-edge models relying on online learning modes, which allow a system to adapt to novel and highly dynamic operational environments. This offers significant tactical and operational gains and makes these systems quite appealing to defence institutions (Payne 2021, 57–79).

When assessed for its ethical implications, the lack of predictability of AI systems poses serious ethical challenges concerning control, the attribution of moral responsibility and accountability, and human autonomy (Samuel 1960; Wiener 1960; Taddeo and Floridi 2018). These challenges are exacerbated when AI systems are deployed for adversarial and kinetic uses, where unique challenges related to the use of force and the presence of adversarial behaviour compound problems already present in civilian uses of AI. As Holland Michel states in a report for the United Nations Institute for Disarmament Research,

> "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. This is because by design, such systems will navigate situations that the operators cannot anticipate" (Holland Michel 2020, 5- emphasis added)

Holland Michel provides the example of a fully autonomous drone mapping 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 will encounter inside the

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

tunnels, and therefore they will not know in advance what exact actions the drone will take" (Holland Michel 2020, 5).

That autonomous systems are intended *by design* to operate in specific environments unanticipated by either operator or programmer is what renders these systems unpredictable in a way which exceeds automatic weapons. This point is echoed in a report by the ICRC (ICRC 2019, 11) with respect to 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."

Operational unpredictability thus pertains to the interaction of the autonomous system with its environment. The nature of this interaction exists at the confluence of a mass of variables, including but not limited to, duration of deployment; environmental complexity; opponent behaviour; the number of systems interacting; the function being performed by the system; along with the understanding the operator has of both the system and the environment of its deployment. It is not feasible to map all these variables nor their manifold interactions and so it will always be difficult to anticipate the effects of LAWS with certainty even when the system is predictable in a technical sense.

2.1. Predictability and Risk of LAWS – The ICRC Position

The unpredictability of LAWS poses problems with respect to Article 36 of Additional Protocol I to the Geneva Conventions on weapons review. As Goussac writes,

> "Foreseeing effects may become increasingly difficult as weapon systems become more complex or are given more freedom of action in their tasks, and therefore become less predictable."

She goes on to explain that

> "The ability to carry out a legal review of a weapon system that utilizes AI entails a full understanding of a weapon's capabilities and foreseeing its effects, notably through verification and testing" (Goussac 2019).

Additionally, the lack of predictability of LAWS poses questions about whether they can be used in a way which adheres with International Humanitarian Law (IHL) and its principles of necessity,

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

proportionality, and distinction. For instance, it will be difficult to use LAWS in a way which respects the principle of distinction between combatants and non-combatants if the system cannot be relied upon to engage with the intended targets only. As the ICRC reports, given the complexity of modern combat situations, particularly urban operations,

> "…compliance with the principle of distinction and rules protecting combatants *hors de combat* already presents formidable challenges. The introduction of AWS to target persons can only increase these challenges…[I]t is difficult to envisage realistic combat situations where AWS use against persons would not pose a significant risk to IHL violations" (ICRC 2021, 9).

Indeed, the principle of distinction is known in international law as *jus cogens*, i.e. it is a fundamental principle of international law from which no derogation is permitted. Given its absolute character, the challenges presented by unpredictability in respecting the principle of distinction have motivated calls for either a moratorium or an outright ban on the use of LAWS in conflict. The position recently adopted by the ICRC on LAWS notes that LAWS bring

> "risks of harm for those affected by armed conflict, both civilians and combatants" because of the difficulties in "anticipating and limiting their effects" (ICRC 2021, 2).

On this basis, the ICRC identifies and suggests three limits on the use of LAWS. First, referring to the principle of distinction, the ICRC recommends that,

> "Unpredictable autonomous weapon systems should be expressly ruled out, notably because of their indiscriminate effects. This would best be achieved with a prohibition on autonomous weapon systems that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained" (ICRC 2021, 2).

Second, on the imperative to uphold "international humanitarian law rules for the protection of civilians and combatants *hors de combat*," the ICRC recommends that "use of autonomous weapon systems to target human beings should be ruled out." This is to be achieved through a prohibition on LAWS "that are designed or used to apply force against persons." Third, the ICRC recommends that LAWS which are not prohibited should be regulated. This includes limiting the type of targets to those "objects that are military objectives by nature"; limits on the duration and scale of use so as to "enable human judgement and control in relation to a specific attack"; limits to uses in situations where "civilians or civilian objects" are absent; and allowing "effective human supervision" of the system.

We agree with the ICRC that the unpredictability of LAWS means they cannot be used in a way that respects the principle of distinction, and that, therefore, LAWS should not be used against

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

human agents. However, we depart from that position as we argue that (i) the ICRC does not provide a strong enough case for its recommendation and that (ii) the recommendation is too permissive.

To support its recommendation to ban LAWS, the ICRC states that LAWS bring "risks of harm for those affected by armed conflict." This is true of all weapons systems. The question is whether LAWS introduce *unacceptable* risk to those affected by armed conflict; in particular, whether LAWS introduce unacceptable risk to non-combatants. If LAWS introduces unacceptable risk to non-combatants, then its use against human agents is unjustifiable. Unfortunately, the ICRC's invocation of the principle of distinction alone, without fully defining that principle, is not sufficient criteria for determining the unacceptability – or not – of this risk. Despite its centrality to IHL, the principle of distinction has numerous interpretations (Bica 1998; Kasher 2007). Minimally, it states that combatants should not target *intentionally* non-combatants. It thereby allows the *unintentional* (though foreseeable) targeting of non-combatants. As Orend writes, the principle of distinction "does not make it illegal for civilians to die in wartime*" per se* (Orend 2019, 112). However, Walzer argues – and we agree with this argument – that the principle of distinction must be supplemented with the obligation of due care (Walzer 1977, 151–59). We argue in the next section that an interpretation of the principle of distinction which includes the obligation of due care, as drawn from the ethical framework of Just War Theory, sheds a much clearer light on the reasons for prohibiting the use LAWS.

The permissiveness of the ICRC's position is in part established by its characterisation of LAWS. In the first place, the ICRC divides LAWS into two categories: LAWS which are *unpredictable* by design (or use) and those which are *predictable*. This dichotomy requires unpacking. If limited to the technical aspects of the LAWS this distinction is in principle correct. As discussed above, there are models of LAWS such as offline models which, considered in isolation, do not necessarily present an issue for predictability. However, all autonomous systems, whether or not predictable by design, will show operational unpredictability through the interaction of the system with its specific context of deployment. We therefore suggest that any analysis focusing on the moral (and also legal) permissibility of LAWS must begin from the recognition that operational unpredictability is something pertaining to the deployment of all types of LAWS. Thereafter, rather than differentiating between predictable and unpredictable systems, any recommendation on LAWS should begin by specifying which steps in the process of exerting force may be within the remit of the LAWS.

The ICRC recommends that the "use of autonomous weapon systems to *target* human beings should be ruled out" (ICRC 2021, 2 - emphasis added). As the UNIDIR outlines (Ekelhof and Persi

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Paoli 2020, 4) , 'targeting' can include a range of tasks such as finding, tracking, and engaging the target before the point of applying force. However, the ICRC's recommendation takes 'targeting' to be only the application of force. Thus, the ICRC explicitly prohibits at most the use of LAWS to apply force against humans. To comply with the obligation of due care, the remit of the LAWS would have to be considerably smaller than that recommended by the ICRC. Establishing the status of a human agent, whether they are liable or non-liable to attack – i.e. the 'right' target – is a difficult task. Determining the liability of an agent to attack, therefore, requires careful assessment prior to the application of force. This is especially important to reduce the risk posed to non-combatants under the obligation of due care. What has not been explicitly forbidden by the ICRC position has been tacitly permitted. And permitting all targeting functions except the application of force to fall under the remit of LAWS will not afford the human operator sufficient opportunity to develop their situational awareness about the nature of the target, infringing the obligation of due care. Thus, the key question is: what is the level of situational awareness afforded to the combatant by the LAWS at a specified targeting function? Is that situational awareness in accord with the obligation of due care (i.e. the reduction of non-combatant risk) given the cognitive load entailed in applying the principle of distinction?

We develop these two objections to the ICRC recommendation further in sections 3 and 4, respectively. First, we unpack the obligation of due care and its relationship to risk and the unpredictability of LAWS. By drawing on the obligation of due care we are drawing on Just War theory. In doing so, we recognise that the challenges which LAWS pose to ethical use are not simply a question of technical capacities, or of developing more refined models. Rather, there are inherent properties to LAWS (as defined in Taddeo and Blanchard, forthcoming) which challenge Just War theory at a fundamental normative and conceptual level.

3. Distinction & Due Care

The principle of distinction requires that parties to an armed conflict distinguish between military objects and civilian objects and direct their actions only towards the former. The distinction is underpinned by liability and non-liability to attack. A combatant is liable to attack for they have abrogated their right not to be attacked. Because combatant liability is established by convention its moral basis is disputed in Just War theory. However, the most widely accepted account is that combatants are legitimate targets because, through their capacity to harm, they pose a threat. Non-combatants are non-liable to attack. Because they have no business in war, they are said to have 'immunity' from the effects of war. The doctrine of non-combatant immunity is absolute in Just War

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

theory: "They [non-combatants] can *never* be the objects or the targets of military activity" (Walzer 1977, 151, emphasis added).

Note however that this principle prohibits combatants from *intentionally* harming non-combatants. The principle of distinction permits the unintentional but foreseeable harming of non-combatants, if that harm is proportionate to the goals the attack is intended to achieve. The permissibility of unintentionally but foreseeably harming non-combatants is called the doctrine of double effect. The doctrine captures the moral intuition that

> "it is permissible to cause a harm as a side effect (or 'double effect') of bringing about a good result even though it would not be permissible to cause such a harm as a means to bringing about the same good end" (McIntyre 2004).

In *Just and Unjust Wars*, Walzer outlined the four conditions of the doctrine of double effect in war, all of which had to hold for the killing of non-combatants to be permissible. The four conditions are:

1) The act is good in itself or at least indifferent, which means…that it is a legitimate act of war.
2) The direct effect is morally acceptable – the destruction of military supplies, for example, or the killing of enemy soldiers.
3) The intention of the actor is good, that is, he aims only at the acceptable effect; the evil effect is not one of his ends, nor is it a means to his ends.
4) The good effect is sufficiently good to compensate for allowing the evil effect; it must be justifiable under… [the] proportionality rule. (Walzer 1977, 153)

These conditions give a minimal account of the principle of distinction. Applied to the use of LAWS, this account does not forbid the use of LAWS. Even given the unpredictability of LAWS, it remains possible that an operator employs LAWS in a way that is permissible under these conditions. In other words, it is possible for an operator to employ LAWS with good intention, aiming at an acceptable effect. It may be that the unpredictably of LAWS meant it was *foreseeable* that harm would come to non-combatants, but if that good effect (i.e. the military objective) was nevertheless "sufficiently good to compensate for allowing the evil effect" (Walzer 1977, 153) under the proportionality rule, then that foreseeable harm is permissible.

For example, consider a building containing one enemy combatant and two non-combatants. An operator chooses to deploy a LAWS to the building with the intention of killing the enemy combatant. Given the unpredictability of the LAWS, it is foreseeable that harm will come to the non-combatants in the building. However, the enemy combatant is a high-value target and their death may bring considerable military advantages, in turn this may hasten on the end of the war and spare many

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

lives. In addition, the operator believes there is a good chance of success in killing the target if the LAWS is deployed. In which case, the good effect of using the LAWS would outweigh the evil effect of bringing harm to non-combatants. The deployment of LAWS, though unpredictable, would therefore be justifiable under the principle of distinction.

However, the minimalist account fails to consider the principle of distinction with the obligation of due care. This obligation, a key element of distinction in Just War Theory, is crucial when considering LAWS, as it defines the risks and obligations of combatants in using lethal force. 'Due care' states that combatants are obliged to accept greater risks to themselves to ensure that they hit only the right target in order to diminish the risks to non-combatants (Orend 2001, 12–13).[5] The nature of due care has not been given full articulation in Just War theory, but it is widely held to be a central tenet of *jus in bello* conduct. As McMahan writes,

> "…the dominant view within the just war tradition…is that when combatants must choose between imposing a certain risk on civilians as a side effect of their action and accepting an even greater risk to themselves, they must, up to a certain point, accept the greater risk" (McMahan 2010, 344).

Thus, at heart of 'due care' is the question of risk allocation, "how to resolve the trade-offs between 'force protection' and minimizing the harm one's forces cause to civilians" (McMahan 2010, 343). We shall detail the nature of this trade-off in section 3.1, along with the threshold at which combatants cease to be obliged to shoulder additional risk. Here, we wish to highlight that the obligation of due care is seen as an extension of the principle of distinction because it extends the moral force of distinction beyond the permissibility of intentional killing (Avishai Margalit and Michael Walzer 2009). As McMahan writes,

> "Even harms that are side effects of permissible action should, when possible, be suffered by combatants rather than non-combatants" (McMahan 2010, 350).

This is so because due care partakes in the same moral bases as the principle of distinction itself. Combatants are liable to harm in war because of their capacity to injure. It follows that if a combatant is liable to harm through intentional attack, then liability should also encompass liability to harm from the side effects of military action. As such it is the role of a combatant to accept greater risk and to avoid imposing them on those who are non-liable.

[5] 'Due care' has been codified in the Law of Armed Conflict as 'precautions in attack' (see: Ministry of Defence 2011, 81–88).

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

> "This is what each side should say to its soldiers...By wearing a uniform, you take on yourself a risk that is borne only by those who have been trained to injure others (and to protect themselves). You should not shift this risk onto those who haven't been trained, who lack the capacity to injure…" (Avishai Margalit and Michael Walzer 2009).

Moreover, due care strongly affirms the 'good intention' highlighted by the doctrine of double effect through the demonstration of restraint in warfare.[6] In effect, the exercise of due care makes good intention – a matter of the combatant's private conscience – publicly accessible – i.e. demonstrable to others. Demonstrating restraint, for instance, might entail

> "…that soldiers use only certain kinds of weapons (e.g. 'smart' bombs, laser-guided cruise missiles), move in more closely on the targets (e.g. flying lower on a bombing raid), gather and analyze intelligence on the precise nature of suspected targets, perhaps provide some kind of advance warning to nearby civilians, and certainly plan the tactic in advance with an eye towards minimizing civilian casualties" (Orend 2001, 13).

The high-altitude bombing campaign by NATO forces in Kosovo, for instance, has been criticized for violating the obligation of due care. Whilst high-altitude bombing

> "enabled NATO combatants to fight with little risk to themselves, it also prevented NATO from being able to locate, identify, and attack its targets with the degree of accuracy and discrimination"

which other means, such as ground troops, could have afforded (McMahan 2010, 342). Walzer has even suggested that when combatants take fire from a building the onus is on those combatants to approach the building to try to get close enough to see who is inside and who are the proper targets of their return fire.

Due care therefore requires that the foreseeable evil of military action be reduced as far as possible. Thus, it modifies the doctrine of double effect, making it more restrictive and, in turn, the principle of distinction more restrictive. Under this modification the principle of distinction is satisfied if

> "[t]he intention of the actor is good, that is, he aims narrowly at the acceptable effect; the evil effect is not one of his ends, nor is it a means to his ends, and, aware of the evil involved, he seeks to minimize it, accepting costs to himself" (Walzer 1977, 155).

[6] See condition (3), page 17.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

3.1. The Allocation of Risk, LAWS, and Necessity

A more restrictive doctrine of double effect and principle of distinction pose a question as to risk allocation and the threshold at which combatants cease to be obliged to shoulder additional risk. Combatants are not required to accept additional risk should that risk 'doom' a justified military objective or "make it so costly that it could not be repeated." (Walzer 1977, 157).[7] This threshold cannot be calculated in abstract. It is a contextual matter requiring situational judgement, for the threshold will vary according to "the nature of the target, the urgency of the moment, the available technology and so on" (Walzer, 156). Indeed, the existence of the available technology in the given context points to the relative nature of the calculation required in establishing the extent of due care: is there an alternative means which reduces the risk posed to non-combatants without jeopardising the given military objective?

As described in the previous section, the unpredictability of LAWS means that it is more difficult to anticipate the outcomes of using LAWS relative to other – say, automatic – weapon systems. The combatant that uses LAWS, employing a system which they know to be unpredictable in its outcomes, cannot be said to have 'intended' to "aim narrowly at the acceptable effect." The use of a weapon with unpredictable outcomes therefore infringes upon the obligation of due care; the question is whether it does so unnecessarily. That is to ask whether there a plausible military objective which would necessitate the use of LAWS against human agents despite the infringement of due care that LAWS entail. When viewed through Just War theory, it is conceivable but implausible that necessity permits the use of LAWS in contravention of due care. It is conceivable, because the Just War doctrine of the 'supreme emergency' justifies the use of any means if the stakes are high enough to warrant them, such as survival (Walzer 1977, 251–68). As Walzer writes, the supreme emergency entails,

> "…fear beyond the ordinary fearfulness (and the frantic opportunism) of war, and a danger to which that fear corresponds, and that this fear and danger may well require exactly those measures that the war convention bars" (Walzer 1977, 251)

This is implausible, however, for two reasons. First, such circumstances seldom exist. Supreme emergencies are rare in history, and even when supreme emergencies may occur necessity remains a relative calculation. So long as there remain other weapon systems that can meet the emergency, but have comparatively lower risks to non-combatants than LAWS, it would be injudicious to use LAWS.

[7] By *justified* military objective is meant that the objective conforms with the other principles of *jus in bello* conduct, such as the principle of necessity and proportionality.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Second, the operational advantages which are often ascribed to LAWS far exceed the capabilities of humans, thereby lowering the threshold for the necessity of using them against human combatants. As the US Air Force reports, the advantages offered by autonomous systems can be for their operating within "extreme performance parameters" such as 'hypersonic flight' (USAF 2009, 41). In such cases, their use is thus necessitated at "a tempo faster than humans can possibly achieve…." (Etzioni and Etzioni 2017, 72). Likewise, the United Kingdom has highlighted that the advantages offered by autonomous systems will be for their speed of decision-making because the "increasing speed, confusion and information overload of modern war may make human response inadequate…" (Ministry of Defence 2011a, nn. 5–10). The advantages ascribed to LAWS in these examples are for their capacity to operate in *excess* of what is required to meet a threat from human combatants. This reduces their likelihood of being an appropriate response to a human threat comparative to automatic or conventional weapons.

Here, we do not mean that their use against combatants is unjustified *per se*. We assume, as with the majority of the Just War theory, the liability to attack of combatants engaged in warfare. There may be reasons which are sufficient for not using LAWS against combatants despite their liability to attack – reasons of dignity in death, for instance, or rights against arbitrary execution which the algorithmic decision to engage a target might present (Heyns 2013; Birnbacher 2016; Heyns 2017; A. Sharkey 2019) – but these are not our concern here. Rather, our claim is that LAWS are not to be used against humans, inclusive of human combatants, because of the unacceptable and unnecessary risk they introduce to those who are non-liable to attack, i.e. non-combatants.

We have so far dealt with LAWS in broad brush-strokes. The complexity of LAWS requires the specification of which steps in the process of exerting force may or may not justifiably be within the remit of LAWS according to the obligation of due care. At the tactical level of mission execution there are a series of tasks which precede the decision to engage the target and exert force, including finding, fixing, and tracking the target. This is so because the categories entailed by the principle of distinction are indeterminate. Their application to specific contexts requires a situational judgement possessed by human agents. The delegation of this judgement to LAWS – even highly refined models – would infringe unnecessarily the obligation of due care. Exploring this claim will also allow us to address an objection that could be raised against our argument so far: that there are conflict environments which are absent of non-combatants and within which the use of LAWS is justifiable.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

4. Context and Liability

The ICRC prohibits at most the use of LAWS to apply force against humans without human intervention. It recommends effective human supervision over the use of LAWS. In doing so, it follows a number of other recommendations that LAWS must have a human in the loop in order to approve the targets selected and presented by the LAWS (Sharkey 2014). The human supervision of the application of force by the LAWS is meant to mitigate its unpredictability by ensuring that the selected targets are the intended ones.

This alone will not be sufficient to guarantee the use of LAWS in a way which complies with the obligation of due care. This is because, as discussed above, predictability is as much a question of the context in which an autonomous system is used as of its technical characteristics. The greater the complexity of the environment in which an autonomous system is used, the more the unpredictability of the system is compounded. As noted by the UNIDIR,

> "…the more complex the operating environment to which a system is deployed, the more likely it is that the system will encounter inputs for which it was not specifically trained or tested or will display new behaviours that have not been previously observed or validated..." (Holland Michel 2020, 7)

As such, mitigating unpredictability is not just a case of human agents supervising the outcomes of the autonomous system, but of developing appropriate understanding of the way the system works and a situational awareness about the environment in which that system is used. As Roff and Moyes writes, the deployment of LAWS should be done in such a way as to

> "...ensure that the human operator has sufficient degrees of situational awareness with regard to the system's operation in its environment, as well as whether the system is functioning within appropriate parameters" (Roff and Moyes 2016, 2).

The complexity of conflict environments requires situational awareness which exceeds that of civilian environments. Adversarial behaviours, greater risk profiles, and operational speeds increase the degree of unpredictability relative to civilian uses of AI. Where situational awareness is most stringently required is in the application of the principle of distinction. As Chengeta writes,

> "many commentators and even fighters themselves have acknowledged the difficulties encountered on the battlefield as far as applying the rule of distinction is concerned" (Chengeta 2016, 85).

The difficulties in applying the principle of distinction arise from the difficulties in determining the practical application of the categories entailed by this principle. Under an orthodox interpretation of

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Just War theory, liability to attack is determined by class membership rather than a moral sense of guilt or innocence. However, the distinction between military actors and civilian actors was drawn up at a time when warfare took place between (at least) two sets of distinctly uniformed soldiers at some geographical distance – i.e. a 'front' – from civilian centres (Nurick 1945).[8]

Today, asymmetric urban insurgencies entailing non-uniformed irregulars has meant that this line between combatant and civilian is much less visually distinct. In turn, this demands a much greater use of situational judgement on the part of the combatant in order to determine the liability (or not) of an actor to attack. Without visual markers such as uniforms or neatly defined battlefield parameters judgements of liability must primarily be based on a judgement of behaviour, and often these behavioural differences are highly nuanced.[9] As Grut writes, it can be

> "….extremely difficult to distinguish a farmer digging a trench from a member of an armed group planting an improvised explosive device […]" (Grut 2013, 11).

Moreover, even aggressive behaviour need not determine liability. A civilian using a weapon for the protection of their family, and in doing so not seeking advantage for any party to the conflict, is not liable to attack.

That liability is established by contextually-dependent behaviour, this is why the claim made by a number of commentators that the use of LAWS against human agents is justifiable in environments absent of non-combatants is misconceived. Desert environments, submarine warfare, and aerial combat are all given as examples of "less complicated" environments which permit LAWS to be used in adherence to the principle of distinction. As such, granted the unpredictability of LAWS, the absence of civilians removes the likelihood of LAWS inadvertently targeting civilians.

However, these near-ideal situations do not hold. The difficulty introduced by such legal categories as 'combatant' versus 'non-combatant', or 'soldier' versus 'civilian', is that they suggest a sense of durability which in the concrete they lack. Within strict parameters, the status of agent liability underpinning these categories can transform on the basis of a given behaviour. It is not a status

[8] This is not to say that civilians were once relieved of the effects of war. Rather, given the clearer distinction between the set of those actors comprising 'combatants' and the set of those comprising 'non-combatants', the line between those liable to attack and those not liable to attack was marked more explicitly.

[9] Revisionist accounts of Just War theory argue that judgements of liability are a function of moral responsibility for a wrongful threat or harm. Wrongful threat of harm under this account is determined by whether (or not) a combatant fights for a just cause (see: Fabre 2009). This departs from the orthodox interpretation which takes liability as a function of class membership. In instances where the orthodox interpretation refers to the 'innocence of noncombatants', 'innocence' is meant in its original Latin sense, i.e. not involved in harming, or helping to harm (Guthrie and Quinlan 2007, 14). This indicates how behaviour underlays the application of distinction under the orthodox interpretation. We choose not to use a revisionist account of liability in this article for both its divergence from IHL and for the greater practical difficulties of application it introduces over the orthodox interpretation.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

intrinsic to the agents themselves. This is why a similar but alternative claim that LAWS be used only to target "objects that are military objectives by nature" (ICRC 2021, 2) also fails. It is context which determines whether an actor or object is a military objective. For instance, a combatant who has either surrendered or been incapacitated through injury is in that moment rendered *hors de combat* and non-liable to attack. One might reply that there are artifacts such as tanks which are objectively means of war, constituting essentially military targets irrespective of context of use. But this is not the case. A case in point is the so-called 'Highway of Death' in first Gulf War (Mueller 1995). On the night of 26th of February 1991 coalition aircraft attacked and destroyed a column of hundreds of Iraqi manned military vehicles. At the time the controversy of the incident rested on the dispute over whether Iraqi forces were surrendering or retreating to regroup. If the former, then the attacks launched by the coalition forces were potentially in violation of the Geneva Convention (Hersh 2000). The existence of this dispute testifies to fact that the notion of objects which are military objectives by nature is misconceived.

Thus, determining the liability or non-liability to attack within a specific context requires a subjective assessment based on a nuanced situational awareness of that context. Judgements preceding the engagement of a target will impact upon the use (or not) of a weapon system. Determining the liability of an agent to attack therefore requires careful assessment by the system operator prior to the decision to apply force. As noted above, due care requires both that the combatant "aims narrowly at the acceptable effect" and that the combatant seeks to minimize risks posed to non-combatants by "accepting costs to himself" (Walzer 1977, 155). In order to minimize the risks to non-combatant in this way, the onus is on combatants to exercise their judgement to determine the liability of the target. This liability, as described, is not a fixed visual marker but manifested through behaviour. The combatant would therefore have to have an account of a given target's behaviour and adequate situational awareness.

Thus, leaving to the remit of LAWS all but the application of force deprives the human operator of the opportunity for determining that liability.[10] Put another way, a human operator who has not sought to use situational judgement to determine the liability of the target has not sought to

[10] It might be argued that situations whereby all targeting functions *except* the application of force are retained by the human operator, with application delegated to the LAWS, would be compatible with the position we set out here. This would include, for instance, situations where a human operator 'paints' a target with a laser designator system, but the autonomous system is left to determine the exact timing and trajectory of the weapon. However, we discount this sort of situation from our position, since it does not fall under our definition of LAWS.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

diminish – as far as possible – the unpredictability of outcomes associated with system-environment interaction. They have therefore not acted in a way compliant with the obligation of due care.

5. Conclusion

In this article we considered whether the use of LAWS to identify, select, and engage targets could be done in a way which complies with the principle of distinction. Under a fuller account of the principle of distinction, one which includes the obligation of due care, we found that because of their operational unpredictability LAWS (as defined in Taddeo and Blanchard 2021) cannot comply with this principle when used to autonomously target human beings. The obligation of due care entails a trade-off between the risks allocated to combatants and non-combatants. It obliges combatants to accept a greater degree of risk, and to reduce the risk to non-combatants, in order to ensure that they hit the 'proper' target. The challenges involved in anticipating the outcomes of LAWS, as defined in section 2, unacceptably skews that risk towards non-combatants. As such, we found that LAWS cannot be used justifiably against human agents. Moreover, we stressed that given the contextual factors of conflict, it is not enough simply to exclude the task of applying force from the remit of LAWS. Rather, the remit of LAWS must be strictly bounded in order to allow the operator the appropriate level of situational awareness for applying the principle of distinction.

It is worth stressing that our analysis is a comparative one, the obligation of due care imposes a choice of means of warfare that pose lower risks to non-combatants. To summarise, if LAWS as we have characterised them in section 1 were to be used when more predictable means are available, then it would not fulfil the duty defined by the obligation of due care. However, if in theory LAWS were more predictable than other possible means, or if it were possible to define technical measures to mitigate the risks posed by their unpredictability, then these weapon systems would not fall under our analysis.

Finally, it is important to stress that the use of AI to support or perform adversarial and kinetic operations poses serious ethical challenges as well as profound changes to the character of war as those technologies evolve (Blanchard and Taddeo 2022). It is possible that new instantiations of AI may address some of the ethical challenges central to the current debate, e.g. lack of transparency, and that new challenges may emerge that we cannot foresee today. This is why it is crucial that research on the ethical implications of the use of AI in defence continues and is developed in tandem with work focusing on the nature of war waging.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

References

Arkin, Ronald. 2018. 'Lethal Autonomous Systems and the Plight of the Non-Combatant'. In *The Political Economy of Robots*, 317–326. Springer.

Arkin, Ronald C. 2009. 'Ethical Robots in Warfare'. *IEEE Technology and Society Magazine* 28 (1): 30–33.

Avishai Margalit, and Michael Walzer. 2009. 'Israel: Civilians & Combatants'. *The New York Review of Books*, May.

Bica, Camillo C. 1998. 'Interpreting Just War Theory's Jus in Bello Criterion of Discrimination'. *Public Affairs Quarterly* 12 (2): 157–168.

Birnbacher, Dieter. 2016. 'Are Autonomous Weapons Systems a Threat to Human Dignity?' In *Autonomous Weapons Systems: Law, Ethics, Policy*, edited by Claus Kreβ, Hin-Yan Liu, Nehal Bhuta, Robin Geiβ, and Susanne Beck, 105–21. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781316597873.005.

Blanchard, Alexander, and Mariarosaria Taddeo. Forthcoming. 'Jus in Bello Necessity, the Requirement of Minimal Force, and Autonomous Weapon Systems'.

———. 2022. 'Autonomous Weapon Systems and Jus Ad Bellum'. *AI & SOCIETY*, March. https://doi.org/10.1007/s00146-022-01425-y.

Boulanin, Vincent, Moa Peldán Carlsson, Netta Goussac, and Davison Davidson. 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.

Chengeta, Thompson. 2016. 'Measuring Autonomous Weapon Systems against International Humanitarian Law Rules'. *Journal of Law & Cyber Warfare* 5 (1): 66–146.

Collopy, Paul, Valerie Sitterle, and Jennifer Petrillo. 2020. 'Validation Testing of Autonomous Learning Systems'. *INSIGHT* 23 (1): 48–51. https://doi.org/10.1002/inst.12285.

Davison, Neil. 2009. *'Non-Lethal' Weapons*. Palgrave Macmillan.

DIB. 2020. '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.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Ekelhof, Merel, and Giacomo Persi Paoli. 2020. 'The Human Element in Decisions About the Use of Force'. Geneva: United Nations Office of Disarmament Affairs.

Etzioni, Amitai, and Oren Etzioni. 2017. 'Pros and Cons of Autonomous Weapons Systems'. *Military Review*, 10.

Fabre, Cécile. 2009. 'Guns, Food, and Liability to Attack in War'. *Ethics* 120 (1): 36–63. https://doi.org/10.1086/649218.

Goussac, Netta. 2019. 'Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting'. ICRC. 18 April 2019. https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/.

Grut, Chantal. 2013. 'The Challenge of Autonomous Lethal Robotics to International Humanitarian Law'. *Journal of Conflict and Security Law* 18 (1): 5–23.

Guthrie, Charles, and Michael Quinlan. 2007. *Just War. The Just War Tradition: Ethics in Modern Warfare*. London: Bloomsbury.

Heyns, Christof. 2013. 'Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions'. United Nations General Assembly. https://doi.org/10.1163/2210-7975_HRD-9970-2016149.

———. 2017. 'Autonomous Weapons in Armed Conflict and the Right to a Dignified Life: An African Perspective'. *South African Journal on Human Rights* 33 (1): 46–71.

Holland Michel, Arthur. 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.

Horowitz, Michael C. 2019. 'When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability'. *Journal of Strategic Studies* 42 (6): 764–88. https://doi.org/10.1080/01402390.2019.1621174.

ICRC. 2021. 'ICRC Position on Autonomous Weapon Systems & Background Paper'. Geneva: International Committee of the Red Cross.

International Committee of the Red Cross. 2019. '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.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Jeffrey Dastin, and Paresh Dave. 2021. 'U.S. Commission Cites "moral Imperative" to Explore AI Weapons'. *Reuters*, 26 January 2021. https://www.reuters.com/article/us-usa-military-ai-idUSKBN29V2M0.

Kasher, Asa. 2007. 'The Principle of Distinction'. *Journal of Military Ethics* 6 (2): 152–167.

Marchant, Gary E., Braden Allenby, Ronald Arkin, and Edward T. Barrett. 2011. 'International Governance of Autonomous Military Robots'. *Columbia Science and Technology Law Review* 12: 272–316.

Matthias, Andreas. 2004. 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata'. *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

McIntyre, Alison. 2004. 'Doctrine of Double Effect', July. https://stanford.library.sydney.edu.au/entries/double-effect/.

McMahan, Jeff. 2010. 'The Just Distribution of Harm Between Combatants and Noncombatants: The Just Distribution of Harm Between Combatants and Noncombatants'. *Philosophy & Public Affairs* 38 (4): 342–79. https://doi.org/10.1111/j.1088-4963.2010.01196.x.

Ministry of Defence. 2011a. 'Joint Doctrine Note 2/11 The UK Approach to Unmanned Aircraft Systems (Supra Note 5)'. JDN 2-11. Swindon: UK Ministry of Defence, Development, Concepts and Doctrine Centre.

———. 2011b. 'Joint Service Manual of The Law of Armed Conflict (JSP 383)'. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf.

Mueller, John. 1995. 'The Perfect Enemy: Assessing the Gulf War'. *Security Studies* 5 (1): 77–117. https://doi.org/10.1080/09636419508429253.

Nurick, Lester. 1945. 'The Distinction between Combatant and Noncombatant in the Law of War'. *The American Journal of International Law* 39 (4): 680–97. https://doi.org/10.2307/2193409.

Orend, Brian. 2001. 'Just and Lawful Conduct in War: Reflections on Michael Walzer'. *Law and Philosophy* 20 (1): 1–30. https://doi.org/10.2307/3505049.

———. 2019. *War and Political Theory*. Cambridge: Polity.

Payne, Kenneth. 2021. *I, Warbot: The Dawn of Artificially Intelligent Conflict*. London: Hurst & Company.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

Roff, Heather M., and Richard Moyes. 2016. 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons'. In *Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, Geneva, Switzerland*.

Samuel, Arthur L. 1960. 'Some Moral and Technical Consequences of Automation--A Refutation'. *Science* 132 (3429): 741–42. https://doi.org/10.1126/science.132.3429.741.

Seymour Hersh. 2000. 'Overwhelming Force: What Happened in the Final Days of the Gulf War?' *The New Yorker*, 22 May 2000.

Sharkey, Amanda. 2019. 'Autonomous Weapons Systems, Killer Robots and Human Dignity'. *Ethics and Information Technology* 21 (2): 75–87. https://doi.org/10.1007/s10676-018-9494-0.

Sharkey, Noel. 2014. 'Towards a Principle for the Human Supervisory Control of Robot Weapons'. *Politica & Societa* 3 (2): 305–324.

Taddeo, Mariarosaria. 2017. 'Trusting Digital Technologies Correctly'. *Minds and Machines* 27 (4): 565–68. https://doi.org/10.1007/s11023-017-9450-5.

———. 2019. 'Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity'. *Minds and Machines* 29 (2): 187–91. https://doi.org/10.1007/s11023-019-09504-8.

Taddeo, Mariarosaria, and Alexander Blanchard. Forthcoming. 'A Comparative Analysis of the Definitions of Autonomous Weapons Systems'.

———. Forthcoming. 'Ascribing Moral Responsibility for the Actions of Autonomous Weapons Systems (Forthcoming)'.

Taddeo, Mariarosaria, and Luciano Floridi. 2018. 'How AI Can Be a Force for Good': *Science* 361 (6404): 751–52. https://doi.org/10.1126/science.aat5991.

Taddeo, Mariarosaria, Tom McCutcheon, and Luciano Floridi. 2019. 'Trusting Artificial Intelligence in Cybersecurity Is a Double-Edged Sword'. *Nature Machine Intelligence* 1 (12): 557–60. https://doi.org/10.1038/s42256-019-0109-1.

Taddeo, Mariarosaria, David McNeish, Alexander Blanchard, and Elizabeth Edgar. 2021. 'Ethical Principles for Artificial Intelligence in National Defence'. *Philosophy & Technology*, October. https://doi.org/10.1007/s13347-021-00482-3.

UN CCW GGE. 2021. 'Chairperson's Summary'. In *Convention on Prohibition or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects*. Vol. CCW/GGE.1/2021/WP.7. Geneva: United Nations Office of Disarmament Affairs. https://documents.unoda.org/wp-content/uploads/2020/07/CCW_GGE1_2020_WP_7-ADVANCE.pdf.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394

US Department of Defense. 2012. 'DoD Directive 3000.09 on Autonomy in Weapon Systems'. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

USAF. 2009. 'United States Air Force: Unmanned Aircraft Systems Flight Plan 2009-2047'. Washington DC: United States Air Force.

Walzer, Michael. 1977. *Just and Unjust Wars: A Moral Argument with Historical Illustrations*. New York: Basic Books.

Wiener, N. 1960. 'Some Moral and Technical Consequences of Automation'. *Science* 131 (3410): 1355–58. https://doi.org/10.1126/science.131.3410.1355.

Preprint

Electronic copy available at: https://ssrn.com/abstract=4099394