The Doctrine of Double Effect & Lethal Autonomous Weapon Systems

Alexander Blanchard,[1] Luciano Floridi, [2, 3] Mariarosaria Taddeo[1, 2]

[1] The Alan Turing Institute, London, UK

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

[3] Department of Legal Studies, University of Bologna, Italy

Abstract

Within Just War Theory, the Doctrine of Double Effect (DDE) modifies the principle of distinction by reference to the intent of an act: the unintentional though foreseeable killing of noncombatants is morally permissible (providing a proportionality clause is met), and the intentional killing of noncombatants is morally impermissible. One concern is that the development of Lethal Autonomous Weapon Systems (LAWS) has superseded DDE because of the separation they introduce between the agent with intention – the human operator – and the agent who targets – the LAWS. As a result, DDE may be incapable of capturing, and thus evaluating, noncombatant deaths resulting from using LAWS. In this article, we address this concern by proposing a revised account of DDE to address cases of noncombatant harm caused by LAWS. We argue that when LAWS cause harm to noncombatants, a distinctive moral wrong occurs because that harm is instrumental to LAWS deployment. This wrong is a consequence of the fact that military organisations deploying LAWS involve noncombatants in circumstances useful to the military organisation precisely by way of involving those noncombatants.

Keywords: Artificial Intelligence, Doctrine of Double Effect, Intention, Lethal Autonomous Weapon Systems, Just War Theory.

1. Introduction

A report by the UN Security Council describes the use of Lethal Autonomous Weapons Systems (LAWS) in Libya in the Spring of 2020. It states that a retreating column of the Libyan National Army

was "hunted down and remotely engaged" by forces of the Government of National Accord using "unmanned combat aerial vehicles or the lethal autonomous systems" such as STM *Kargu*-2 […] and other loitering munitions." The report then states that

> "The lethal autonomous weapons systems were programmed to attack targets without requiring data connectivity between the operator and the munition: in effect, a true 'fire, forget and find' capability" (Choudhury et al. 2021, 17).

This represents the first recorded use of LAWS in a combat theatre. It also shows an intensification of the delegation to Artificial Intelligence (AI) systems of military tasks once performed by humans (Taddeo et al. 2021a). On this basis, the use of AI for selecting and engaging a target, "without further intervention by a human operator" (US Department of Defense 2012, 13) is likely to be prevalent across global militaries in the not-too-distant future.[1] Indeed, in the current conflict in Ukraine, Russia is documented as using Iran's Shahed-136 'kamikaze' drone. Deployed as a swarm to overwhelm air defence systems, the Shahed-136 is reported to have an autonomous function which allows it to loiter in a specified area before selecting a target to attack (Ramzy 2022). Autonomous drones provide Russia with an affordable offensive system in a conflict where it has struggled to gain air superiority with so-called 'manned' systems. More generally, there are other important factors driving this trend towards autonomy in warfare: humans have been described as the "weak link in the kill chain" (Krishnan 2009, 34), and AI offers defence agencies speed of decision-making in adversarial environments when communications are impaired or severed (Alwardt and Schörnig 2022).

Ethical analyses will be essential for underpinning reflection on the use of these weapons both to support regulation and to foster ethical behaviour 'over and above' such regulation (Floridi 2018, 4). In warfare, Just War Theory represents "a common language for discussing and debating the rights and wrongs of conflict" (Whetham 2010, 65). It provides the domain-specific ethical framework for evaluating the moral permissibility of using LAWS (Blanchard and Taddeo 2022c). Whilst incapable of addressing *all* ethical issues related to LAWS (see: Heyns 2017; Taddeo and Blanchard 2022b), Just War Theory provides the ethical underpinning of International Humanitarian Law (IHL). Therefore, Just War Theory provides an ethical framework applicable by those actors (defence organisations) at the forefront of developing and using LAWS.

[1] Consensus-based definitions of autonomous weapons systems (AWS) remain elusive, making identifying an AWS a difficult task for international organisations (Wyatt 2020; Taddeo and Blanchard 2022c).The developer of the Kargu-2, identified in the UN report, claims that it is capable of fully autonomous flight, whilst engagement decisions remain the remit of the operator (STM 2020). However, whether an AWS is in autonomous targeting mode is a decision undertaken by, and known only to, the military deploying the AWS.

A growing body of literature has sought to evaluate LAWS according to Just War Theory's *jus in bello*[2] principles, namely: distinction, proportionality, and necessity (Abney 2013; P. Asaro 2008; 2011; Blanchard and Taddeo 2022a; 2022b; Boulanin, Bruun, and Goussac 2021; Chengeta 2016; Demy 2020; Galliott 2017; Homayounnejad 2018; Simpson and Müller 2016). The applicability of the principle of distinction (a central tenet of *jus in bello*) to LAWS has attracted much attention. The principle prescribes that LAWS ought to be used respecting the difference between combatants and noncombatants, and the immunity from attack granted to the latter.

In this article, we extend the focus on harm to noncombatants by considering whether the Doctrine of Double Effect (DDE) remains a viable doctrine for assessing harm caused to noncombatants by LAWS. Under DDE, acts with identical consequential profiles can be either permissible or impermissible, depending on the intention of that act. In focusing on the intention of an act, DDE modifies the principle of distinction by making the *unintentional* though foreseeable harming of noncombatants permissible (providing a proportionality clause is met).

When considering LAWS and DDE, one claim is that LAWS have superseded DDE, leaving the latter incapable of addressing cases of noncombatant deaths resulting from their use (Swiatek 2012; see also: Abney 2013). This is because the use of LAWS separates the agent with *intention* – the human operator – from the agent who *targets* – the LAWS. Thus, when a LAWS kills a noncombatant, that harm can be said to be deliberate but *not* intentional. This, in turn, produces a novel category of noncombatant death: a death which is neither collateral nor intended, effectively 'short-circuiting' DDE (Swiatek 2012, 252).[3]

We argue that DDE does remain valid for assessing uses of LAWS and that its application shows that DDE does not justify noncombatant harm resulting from LAWS targeting. This conclusion depends on what account of DDE one uses. Just War Theory may be a common language for discussing the rights and wrongs of war, but it is a language spoken in many idioms, with DDE itself subject to multiple interpretations. An understanding of DDE that takes intention to be an internal state of mind will be inapplicable to LAWS. This view, which begins with Thomas Aquinas, predominates in the Just War Tradition, and we outline it in section 2. In section 3, we outline why LAWS outpace this Thomist account of DDE. In section 4, we argue that, when LAWS target noncombatants, the distinctive moral wrong of using LAWS is that harm to noncombatants is instrumental to their deployment. This wrong is a consequence of the fact that military organisations

[2] I.e., the principles governing conduct *in* warfare.
[3] Our use of 'collateral' in this article is not an indication that we take this term to be rhetorically unproblematic.

deploying LAWS involve noncombatants in circumstances useful to the military organisation precisely by way of involving those noncombatants. We use Warren Quinn's distinction between direct and indirect harmful agency to develop this claim, demonstrating how it assesses the novel category of harm introduced by LAWS. In section 5, we conclude our analysis.

Two clarifications are needed before delving into our analysis. First, in the rest of this article, we refer to the following definition of autonomous weapon systems (AWS):

> "…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 and Blanchard 2022c, 15).

Under this definition, LAWS is a subset of AWS, namely AWS used with the goal of exerting lethal kinetic force against human beings.

Second, our analysis assumes LAWS will be deployed to combat environments where noncombatants are physically present. The proximity of civilians and military objectives in the current conflict in Ukraine exemplifies a trend true of war more generally. In such environments, we can expect the targeting of noncombatants as a statistical inevitability of the functioning of LAWS. In the future, LAWS may be deployed in theatres far from centres of civilian life, such as for anti-submarine warfare, thereby drastically reducing the risk they pose to noncombatants.[4] This poses no issue for our argument, as we ask, *if* LAWS target noncombatants, how can this harm be morally evaluated under DDE?

2. The Doctrine of Double Effect

The DDE captures the moral intuition that

[4] See for example the autonomous Echo Voyager developed by Boeing (2022). Whilst anti-submarine warfare removes the problem of noncombatant targeting present in urban combat, there nevertheless remains the problem of LAWS targeting combatants that are *sans combat*. We cannot address this problem here but have discussed it in (Blanchard and Taddeo 2022b).

> "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).

The principle was given its earliest expression by Thomas Aquinas in a discussion of self-defence. Killing one's assailant, claimed Aquinas, is justified so long as one did not *intend* to kill them, even though such an effect may have been foreseeable (Boyle 1980). DDE has played a prominent role in Just War Theory,[5] offering a doctrine for permissible political violence (Draper 2016, 125). By distinguishing "intentions and outcomes" in the use of violence, Aquinas "created the space for subsequent jurists to formulate practicable laws of war" (Bellamy 2006, 48).

The most critical function of DDE is permitting exceptions to the *jus in bello* principle of distinction, which provides an absolute prohibition against killing noncombatants, i.e., those with immunity from harm. As Walzer writes, "They [noncombatants] can never be the objects or the targets of military activity" (Walzer 1977, 151). The centrality of 'distinction' to the just conduct of war is illustrated in legal codifications of Just War Theory – such as International Humanitarian Law (IHL) – whereby it is deemed *jus cogens* – that is, permitting no derogation. However, this absolute prohibition stands in tension with the fact that, in war, there may be military objectives valuable enough to necessitate the killing of noncombatants (Lee 2004). To resolve this tension, under DDE, noncombatants may be foreseeably killed during a sufficiently justified and necessary military action, but they may not be killed intentionally as the result of that action. The distinction underpinning DDE is illustrated through a discussion of terror bombing and strategic bombing:

> "The terror bomber aims to bring about civilian deaths in order to weaken the resolve of the enemy: when his bombs kill civilians this is a consequence that he intends. The strategic bomber aims at military targets while foreseeing that bombing such targets will cause civilian deaths. When his bombs kill civilians this is a foreseen but unintended consequence of his actions" (McIntyre 2001, 219).

The actions of the strategic bomber are considered permissible because the civilian deaths are incidental to the end the bomber pursues. The noncombatant deaths are what Aquinas called *praeter*

[5] Just War Theory is a contested tradition. In the main it is bifurcated into 'orthodox' and 'revisionist' camps. The Just War Theory we predominantly engage here falls under the 'orthodox' interpretation. This preference is a practical one: the orthodox account of Just War Theory bears the closest relation to the Law of Armed Conflict. However, throughout this article we employ texts and concepts at home in both camps. This follows from a belief that if Just War theorising is to adequately respond to the challenges raised by artificial intelligence, then that theorising must draw on the Just War tradition in its entirety.

intentionem – beside or outside the intention (Boyle 1978, 650). The actions of the terror bomber are considered impermissible because noncombatant deaths are instrumental to her end. The practical importance of this distinction is illustrated through codification of Just War Theory in international law.[6] The Law of Armed Conflict (LOAC) differentiates between harming noncombatants indirectly and directly, with the former being permissible, whilst the latter is considered a war crime (ICRC 2021a).

However, DDE does not provide a blanket permission. DDE operates in conjunction with other *jus in bello* principles like the principles of proportionality and necessity. Taking this into account, Walzer outlined the four conditions of DDE, all of which must be satisfied for the killing of noncombatants to be permissible:

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).

The third condition is crucial for distinguishing whether the good intent was aimed at the good effect, thereby permitting the evil, unintended, effect. Since we are concerned with the problem of LAWS as it relates to intention, we bracket the other conditions for the rest of this article. It is worth highlighting that there may be many good ethical reasons for not employing LAWS besides those relating to DDE. However, they lay beyond the scope of this article.[7]

3. The challenge of applying DDE to LAWS

Commentators have suggested that characteristics intrinsic to LAWS leave one incapable of using DDE to evaluate the permissibility of harm to noncombatants caused by LAWS (Swiatek 2012; Abney 2013). This is because LAWS separate the agent with intention – the human operator – from the agent who targets – the LAWS. Historically, Just War Theory assumes that these are the same agent, namely,

[6] As Draper (2016, 124) points out, the legal distinction between direct and indirect attack "follows the contours" of DDE and historical scholarship suggests it is "quite likely" that DDE provided the rationale for this distinction.
[7] (Blanchard and Taddeo 2022c; 2022b; 2022b; Taddeo et al. 2021b; Taddeo and Blanchard 2022b)

the human combatant. This is attested to by the third condition of Walzer's description of DDE: "The *intention of the actor is good*, that is, *he aims* only at the acceptable effect" (Walzer 1977, 153 - emphasis added). As noted at the outset, LAWS are "able to identify, select and attack the target without the intervention of another agent" (Taddeo and Blanchard 2022c, 15), as they rely on some form of AI capability. While these capabilities enable LAWS to learn from, and adapt to, the environment in which they operate, they do not endow LAWS with intent. It is this unintentional or 'mindless' agency of AI that poses new challenges (Floridi and Sanders 2004). It is worth stressing that the ethical challenges introduced by LAWS are distinct from existing 'fire and forget' systems, whereby projectiles are self-correcting after launch (Payne 2021, 96). Whilst such weapons may demonstrate degrees of autonomy, the decision to engage a particular target remains within the remit of the human operator – i.e., the agent with intent.

The use of LAWS thus produces a novel category of harm, which escapes the dichotomy between intended and collateral harm: the 'non-target target'. A 'non-target target' is an individual killed or harmed by a LAWS who cannot be described adequately as either the 'direct' or the 'indirect' target of that harm (Swiatek 2012). A result of this is the breakdown of the category of 'collateral damage' relating to LAWS.

Consider a scenario where a human operator uses a LAWS to engage and kill a legitimate military target. The operator deploys the LAWS but instead of targeting the legitimate target, the LAWS targets and kills a noncombatant. Assuming that the human operator sought to act in compliance with the principle of distinction, then the civilian harmed by the LAWS was harmed unintentionally. However, the harm to the noncombatant is not 'collateral' harm. Collateral is harm incidental to the intended harm. 'Incidental' is here situational: harm as the product of a blast radius, say, or as the result of a projectile going off-course. Those targeted by a LAWS do not bear this situational relationship to the intended target, because their being targeted was the product of an action of the LAWS. At a first analysis, it seems that LAWS 'short-circuits' DDE "by pushing beyond its ability to account for complex states of affairs" (Swiatek 2012, 252).

In the following sections, we draw on Quinn's distinction between direct and indirect harmful agency to show that DDE is still valid when considering LAWS and captures well what is unethical about these systems. Before that, section 4 details the technological root of the issue through a discussion of the working of LAWS. In the face of claims that LAWS will act 'more morally' than human combatants (Arkin 2018), it is important to explain why there is a reasonable concern that LAWS may target civilians.

4. The Predictability Problem and Data Deficiencies

The predictability of AI systems indicates the degree to which one can answer the question: *what will an AI system do?* Unpredictable systems are not a new issue. They are common in mathematics and physics, and limits about the ability to predict the outcomes of artificial systems have been proven formally since the 1950s (Rice 1956; Moore 1990; Musiolik and Cheok 2021). Wiener and Samuel debated the predictability of AI systems in a famous exchange in 1960 (Wiener 1960; Samuel 1960). Wiener attributed the lack of predictability to the learning abilities of these systems, noting, "as machines learn they may develop unforeseen strategies at rates that baffle their programmer" (Wiener 1960, 1355). Developments in AI research have proved Wiener correct. Consider, for example, reward hacking, which is reported in current literature as one of the factors that can 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."

The predictability of AI systems is debated at technical and operational levels. Some AI researchers focus on the technical features of a system (International Committee of the Red Cross 2019; Boulanin et al. 2020; DIB 2020), while others consider predictability a function of the system and its context of deployment, i.e. operational predictability (International Committee of the Red Cross 2019; Docherty 2020).

From a technical standpoint, the predictability of an AI system is assessed in terms of the degree of consistency between its past, current, and future behaviours (Holland Michel 2020a). Key aspects monitored here are data and concept shift; 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, Sitterle, and Petrillo 2020; DIB 2020).[8] Predictability also depends on properties such as interpretability, transparency, explainability and trustworthiness (Holland Michel

[8] 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 system to behave as expected even when it is fed with erroneous data) (Heaven 2019; Taddeo, McCutcheon, and Floridi 2019a).

2020a; Rudin, Wang, and Coker 2020) of an AI system, insofar as these facilitate envisaging systems outcomes.

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 2020b, 5).

A large set of variables impacts operational predictability: the technical features of the system, the characteristics of the context of deployment, interactions with other systems, the level to which the operator understands how the system works and, in the security domain, the behaviour of adversaries. These variables may change and interact differently, making it problematic to predict all possible actions that an AI system may perform once deployed in a given environment. In this article, we shall refer to the predictability problem as defined by (Taddeo et al. 2022):

> *"Minimally*, given an ideal scenario where no errors at design and development stages can be assumed or detected, once deployed and as a result of their adapting capabilities, an AI system may still develop autonomously correct (and yet unwanted) outcomes which were not foreseeable at the time of deployment.
>
> *Maximally*, given the multi-faced processes of design, development, and deployment of AI systems, the opaqueness of these systems, their adapting capabilities, and the possible complexities of the environment of deployment, it is not possible to account for all sources of errors and manipulation of a system or for all possible emerging behaviours – whether beneficial or not – of an AI system that these errors may prompt" (Taddeo et al. 2022, 15).

This definition enables us to stress that the predictability problem refers to correct and incorrect outcomes. 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.

This lack of predictability of outcomes is central to the ethical challenges posed by AI systems because predictability is a primary metric by which "humans can measure whether their creations are continuing to function as intended" (Roff and Moyes 2016, 2). In the context of armed conflict, the lack of predictability of effects limits compliance with the principles of distinction, proportionality, and necessity, as well as the attribution of responsibility for harm caused (ICRC 2021b; Blanchard and Taddeo 2022c; 2022b; 2022a; Taddeo and Blanchard 2022a).

As stressed by (Taddeo et al. 2022), data and their acquisition, curation, transformations and storage are one of the critical, root causes of the predictability problem. AI systems require quality data (i.e. relevant, complete, and accurate) to perform as desired (Holland Michel 2021, 1). In any operating environment, be it civilian or military, such data is seldom available. Data can be incomplete, perhaps because of an insufficiently high resolution; data may be discrepant because system inputs fall outside the spectrum of training data; and data may be incorrect, perhaps because sensors are poorly calibrated (Holland Michel 2021, 3–5). In all these cases, data can lead the system to develop new, unforeseen outcomes. In the military context, this unpredictability is exacerbated by the presence of adversarial and evasive behaviour, for example, opponents may attempt to fool LAWS through countermeasures like spoofing, i.e., feeding the system 'poisoned' data by introducing minor perturbations. Research has shown that pixel-level perturbations in digital imagery can lead AI systems to misclassify images with high confidence (Szegedy et al. 2014; Uesato et al. 2018). One AI image classifier model was tricked into mistaking an image of a turtle for a rifle (Athalye et al. 2018; Taddeo, McCutcheon, and Floridi 2019b). As Taddeo et al. (2022, 10) stress,

> "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."

Future technological developments may mitigate the predictability problem in its maximal definition, for example, more robust AI systems, and thus more robust LAWS, may be developed in the future. However, due to the operating conditions in wartime, a technological fix addressing these limitations is unlikely. "Though it is difficult…to anticipate exactly what form autonomous weapons will take in the future," writes Michel Holland, "these systems will always contend with data that are problematic and unpredictable" (Holland Michel 2021, 2). Even if one were to assume that all possible sources of perturbation were removed, the predictability of LAWS would remain problematic due to the autonomy and adapting capabilities (as in the minimal definition of the predictability problem).

A way to address the predictability problem is to consider a trade-off between the military advantages of using LAWS and the related risks that these may develop unforeseen (and unwanted) behaviour. This trade-off implies a risk threshold determining the acceptable probability that the LAWS will not act as foreseen by its designers, developers, and deployers. Herein lies the trade-off resulting from the operational unpredictability of the system confronting any military organisation deploying LAWS:

> "to make it more effective, i.e. to produce more targets, we have to lower our expectation that it's targeting the right people" (Swiatek 2012, 249).

The risk resulting from this approach is to consider "the deliberate targeting of noncombatants as a statistically necessary function of the system" (Swiatek 2012, 241).

5. Revising DDE

The problem with DDE, as Swiatek writes, is that

> "it doesn't appear to be particularly suited to account for greyish areas of intention, any more than it can make space for greyish categories of victims" (Swiatek 2012, 252).

Therefore, the challenge's nub is to formulate an account of DDE capable of capturing, in conceptual terms, the intentional structure entailed by the use of LAWS, as described above, without requiring recourse to the intent of the human operator.

To this end, Swiatek proposes evaluating noncombatant harm by reference to the control exerted by those establishing thresholds for error. He claims that there is a degree of control exercised by decision-makers (e.g., within military organisations) in the deployment of LAWS, according to which we can label the killing of noncombatants as intentional. After all, a complex decision-process is undertaken, the match threshold as a statistical probability is specified based on that decision-process, and noncombatants are killed in line with that specification. Searching for nuance, Swiatek proposes the concept of 'intentional error' to capture the killing of noncombatants by LAWS. He writes that

> "It's precisely this attempt to impose our control over error that opens the possibility to label the killings as intentional. There is a stronger measure of control associated with bringing about an error rate of 'x' than there is with merely foreseeing unwanted outcomes occurring with 'x' probability" (Swiatek 2012, 250).

Swiatek explains that, in the former case, the military organisation exercises control over the noncombatant killings by specifying the rate of miscalculation in a way that constitutes "effective action." Without specifying the threshold of error, there would be no noncombatant harm since the LAWS could not be deployed. The military organisations deploying the LAWS and specifying the threshold of error thereby "bring about" the deaths of the noncombatants (Swiatek 2012, 250). At the same time, in the case of "merely foreseeing unwanted outcomes," military organisations seek to avoid miscalculation (Swiatek 2012, 250). This might be done by undertaking extensive testing of the system, extensive training of those using the system, and implementing Rules of Engagement to guide the use

of those systems. On this distinction between 'bringing about' and 'foreseeing', Swiatek suggests we ground ethical judgements about the use of LAWS. He explains that

> "Establishing a tolerance for error constitutes an additional step in the exercise of control and concomitantly, greater responsibility for the outcomes…Thus we have the qualities generally associated with an intentional act (volition, control, causation, etc.) upon which judgements about propriety may be made, and for which we should rightly be held accountable" (Swiatek 2012, 250).

We disagree with Swiatek's proposal for revising DDE because his distinction does not preserve the moral intuition that undergirds DDE. It is not control exercised in bringing about a state of affairs which distinguishes licit from illicit effects under DDE, but the intent in doing so. Recall the contrasting pair of the strategic bomber and the terror bomber. The two bombers both bring about an equivalent effect. The moral value of each act differs according to the absence (or presence) of intent to harm civilians on the ground. In each case, the degree of control exercised over the death of the civilians is not a pertinent moral factor in the contrasting cases when assessing the permissibility of each act under DDE. Indeed, if in moral judgement there is recourse only to the degree of control exercised over a situation, what is left is an assumption, typically found in anglophone law, that "any outcome foreseen as liable to result from a contemplated action must be regarded as intended" (Gould 2014, 134). Accepting this assumption would vitiate the capacity of Just War Theory to distinguish between permissible acts of war and terroristic violence (McMahan 1994b, 201).

5.1. Direct and Indirect Harmful Agency

If the challenge of applying DDE to LAWS arises from its unsuitability for capturing 'greyish' areas of intention, then it is necessary to depart from an account of DDE that centres on a concept of intention as a state of mind. Indeed, this view of DDE has been criticised for introducing far too permissive attitudes towards acts of violence. For example, as Anscombe writes,

> "… if intention is all important – as it is – in determining the goodness or badness of an action, then on this theory of what intention is, a marvellous way offered itself of making any action lawful. You only have to 'direct your intention' in a suitable way. In practice, this means making a little speech to yourself: 'What I mean to be doing is…'" (Anscombe 1961, 58).

Historically, the Thomist account of DDE arose from a concern for the moral character of those fighting in war (Lang 2016). This had its basis in a theological worldview whereby the salvation of the immortal soul for those waging war was of the utmost importance (Bellamy 2006, 30–48). A theory

of political violence preoccupied with the persistence of the soul in the afterlife will have little to say about those agents, autonomous systems, that have "no soul to damn" (Asaro 2012).

Therefore, capturing the distinctive moral wrong of causing harm to noncombatants using LAWS requires an account which does not pivot on the Thomist criterium of *praeter intentionem*. We argue that noncombatant harm is instrumental for the deployment of LAWS, and this comprises the moral wrong upon which DDE should be reformulated. The distinction that Quinn draws between two types of harmful agency can make this claim perspicuous:

> "…between agency in which harm comes to some victims, at least in part, from the agent's deliberately involving them in something in order to further his purpose precisely by way of their being so involved (agency in which they figure as *intentional objects*) and harmful agency in which either nothing is in that way intended for the victims or what is so intended does not contribute to their harm" (Quinn 1989, 343- emphasis original).

Quinn calls the first kind of agency 'harmful direct agency', and the second 'harmful indirect agency'. According to Quinn's account of DDE, we need "*ceteris paribus*, a stronger case to justify harmful direct agency than to justify equally harmful indirect agency" (Quinn 1989, 344). We depart from Quinn in that his formulation of this distinction unnecessarily retains the language of 'intention'. However, the insight in Quinn's distinction is that DDE can be premised on the distinctive wrong entailed by instrumentalising a person(s) for one's ends, regardless of whether such instrumentalisation is intended (or deliberate, to use Quinn's vocabulary). This has for its rationale the Kantian ideal of human community: the fundamental moral status of persons is as possessing independence of choice over their purposes and the uses that can be made of their bodies. This entails that we are each to be treated as existing for purposes that we can share, and that it is wrong to involve persons in the advancement of purposes they cannot share without their consent. As Quinn writes, those that can

> "be usefully involved in the promotion of a goal only at the cost of something protected by their independent moral rights (such as their life, their bodily integrity, or their freedom) ought, *prima facie*, to serve the goal only voluntarily" (Quinn 1989, 349)

Thus, the benefit of using Quinn's distinction for assessing LAWS is that it enables us to "sidestep" the question of an agent's intent in having caused harm to a noncombatant, whilst retaining the intentional structure pertaining to how the harm was caused. What distinguishes the two forms of harmful agency is whether an agent causes harm to an individual in consequence of having involved that individual in serving the agent's ends.

If we consider the case of the terror bomber under Quinn's distinction, it is not strictly speaking important whether the terror bomber 'undeniably intended' the deaths of the civilians. Of moral significance is that the terror bomber harms civilians as a means of achieving their purpose, and that the civilians did not consent to that harm. This is different when considering the strategic bomber, as the harm that the strategic bomber brings to civilians does not further their purposes. The purpose of strategic bomber is to hit a military target:

> "Of course, he [the strategic bomber] is well aware that his bombs will kill many of them, and perhaps he cannot honestly say that this effect will be 'unintentional' in any standard sense, or that he 'does not mean to' kill them. But he can honestly deny that their involvement in the explosion is anything to his purpose" (Quinn 1989, 342).

The strategic bomber would still (indeed would prefer to) deploy its bombs even if there were no casualties among noncombatants, whereas this would utterly undermine the action of the terror bomber.

This sheds light on the wrongfulness of LAWS harming noncombatants under DDE. It may well be that the military organisation, and the human operator acting on behalf of that organisation, can reasonably deny intending harm to noncombatants when the LAWS they deploy target the latter. Yet, adapting Quinn's distinction, whether those deploying the LAWS did or did not intend harm to noncombatants is a question of lesser importance. What is of greater importance is whether the harm came to noncombatants by way of the military organisation involving them to further its purposes, and whether that purpose was furthered "precisely by way of their being so involved" (Quinn 1989, 343).

Under this view, the harm brought to noncombatants by LAWS, in the manner outlined above, appears to be a case of direct harmful agency. The noncombatants may be harmed precisely by way of their being involved to further the pursuit of a military objective. The unpredictability of LAWS introduces noncombatant targeting as statistically inevitable for the functioning of the system. Without allowing for noncombatants being so involved, the LAWS could not be deployed, and if it cannot be deployed then it cannot be used in pursuit of the desired military objective, i.e., for the given purpose. The distinctive moral wrong is that harm to noncombatants is instrumental to using LAWS.

As noted, Quinn states that we need "*ceteris paribus* a stronger case to justify harmful direct agency than to justify equally harmful indirect agency" (Quinn 1989, 344). This raises the possibility that there might be special rights permitting harmful direct agency. In making this qualification, Quinn appears to have in mind special rights or permissions that might arise in activities distinct from war-

waging – for instance, the bringing about of direct harm to a foetus to ensure the mother's health. It is difficult to conceive of special rights pertaining to the activity of war which would permit direct harm to noncombatants. This is because the categories of noncombatant and combatant are themselves rights-based (Lazar 2017). The combatant is said to be 'liable' to attack because they have surrendered their right to life by posing a lethal threat to others. As Walzer states, "simply by fighting, whatever their private hopes and intentions, they have lost their title to life and liberty" (Walzer 1977, 136). Under the account of Just War Theory proposed by Walzer, combatants surrender these rights irrespective of whether they fight for a just or an unjust cause. Conversely, noncombatants retain their rights, so they cannot be legitimate objects of attack.

It is true that there are alternative accounts of liability to attack in war which uncouple the categories (non)combatant from questions of liability (Lazar 2009). For instance, under the revisionist account, liability to attack arises from blameworthy moral responsibility for a wrongful threat (McMahan 1994a; Otsuka 1994). Since noncombatants can be responsible for a wrongful threat – such as voting for or financially supporting through taxation a belligerent government – this would undermine noncombatant immunity in the way conceived by Walzer. Nevertheless, this remains a right-based conception of liability. It does not move us beyond questions of liability to attack *per se*; it only shifts the boundaries of the persons considered liable to attack according to how rights against harm have been lost. Those responsible for wrongful threats are now regarded as liable, whilst those that are not responsible are non-liable. This does not introduce special permissions; it is not permissible for those non-liable to attack to be the victims of direct harmful agency. It still remains that, as Walzer writes, "a legitimate act of war is one that does not violate the rights of the people against whom it is directed" (Walzer 1977, 135).

Quinn also raises the case of harm brought to persons who "physically get in the way of our otherwise legitimate targets…" (Quinn 1989, 345). Consider a possible scenario whereby LAWS legitimately targets and engages a combatant but where the noncombatant physically enters the line of fire and is harmed. We take such harm to be a case of indirect harmful agency. Such cases fall under the category of collateral damage. This does not make such harm permissible *per se*, as it must respect the principle of proportionality. But it is not a case of direct harmful agency, because in such cases, it is not for the purpose of the military organisation that the harm occurs; neither does the possibility that instances of noncombatant harm caused by LAWS will be indirect harm permit the use of these systems.

Therefore, when LAWS target noncombatants, harming them would be morally wrong insofar as the deployment of LAWS requires as a statistical inevitability noncombatant harm, under the conception of DDE outlined above. LAWS treat noncombatants as a means to an end, something they cannot avoid doing for statistical reasons. Treating noncombatants as a means to an end is unacceptable on the grounds outlined above.

Two implications follow from this conclusion. First is that the moral permissibility of LAWS is not determined solely by calculating their effects. Some commentators have argued that LAWS are morally desirable if their rates of noncombatant targeting are fewer than those of human combatants (Arkin 2009; 2010; 2018; Grut 2013; MacIntosh 2021; Marchant et al. 2011; Riesen 2022; Scholz and Galliott 2021; Umbrello, Torres, and De Bellis 2020). This omits crucial non-consequentialist considerations. To be sure, the consequences of war are important to Just War Theory, but it also gives significant weight to deontological distinctions, like the distinction, discussed above, between intending and foreseeing harm. Our approach has been to reconstruct this distinction on Kantian grounds so to retain its normative structure whilst capturing the technological and philosophical novelties of LAWS. In so doing, we retain what is vital to Just War Theory: the ability to discriminate between different types of evil effects (Hurka 2010, 26). Under DDE, unintended but foreseeable harm to noncombatants by *human* combatants is permissible (providing the proportionality clause is met); whereas noncombatant harm caused by LAWS is never permissible.

Second, our conclusion is based on the current state of the technology of LAWS. If AI systems underpinning LAWS can be made entirely predictable, and rates of noncombatant targeting are thereby reduced to zero, then our argument will be superseded. However, we believe this to be very unlikely. Under the maximal definition described above, the unpredictability of AI is a product of environmental complexities and technical characteristics, and there is no 'fix' for the environmental complexities of wartime.

6. Conclusion

In this article, we have addressed the applicability of the Doctrine of Double Effect (DDE) to Lethal Autonomous Weapon Systems (LAWS). We have shown that military organisations deploying LAWS involve noncombatants in circumstances that are useful to the military organisation precisely by way of involving those noncombatants. Because of the technical features of LAWS, harm to noncombatants is instrumental to the deployment of LAWS. This is a distinctive moral wrong and provides the means for evaluating LAWS under DDE. As modern battlefields are increasingly

populated with human beings and artificial intelligence systems, and as technology used for LAWS matures, developing an account of DDE that can speak with one voice across all types of agents will be of increasing importance. Indeed, the need to retain DDE raises a fundamental question, whether Just War Theory is a suitable normative framework for evaluating and governing contemporary conflict. Commentators claim that AI marks a sea-change in warfare, transforming it in barely conceived ways, rendering frameworks such as Just War Theory incapable of providing ethical guidelines governing war. However, as we have shown above, so long as writers on the ethics of war draw broadly from the Just War tradition, such claims remain unfounded for the time being.

References

Abney, Keith. 2013. 'Autonomous Robots and The Future of Just War Theory'. In *Routledge Handbook of Ethics and War: Just War Theory in The Twentieth-First Century*, edited by Fritz Allhoff, Nicholas G. Evans, and Adam Henschke. London.

Alwardt, Christian, and Niklas Schörnig. 2022. 'A Necessary Step Back?: Recovering the Security Perspective in the Debate on Lethal Autonomy'. *Zeitschrift Für Friedens- Und Konfliktforschung*, February. https://doi.org/10.1007/s42597-021-00067-z.

Anscombe, G. E. M. 1961. 'War and Murder'. In *Nuclear Weapons: A Catholic Response*. London. https://philarchive.org/archive/ANSWAM.

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

———. 2010. 'The Case for Ethical Autonomy in Unmanned Systems'. *Journal of Military Ethics* 9 (4): 332–41.

———. 2018. 'Lethal Autonomous Systems and the Plight of the Non-Combatant'. In *The Political Economy of Robots*, 317–26. Springer.

Asaro, P. M. 2012. 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics'. In *Robot Ethics: The Ethical and Social Implications of Robotics*, edited by P. Lin, K. Abney, and G. A. Bekey, 169–86. MIT Press. https://ieeexplore.ieee.org/document/6733967.

Asaro, Peter. 2008. 'How Just Could a Robot War Be'. *Current Issues in Computing and Philosophy*, 50–64.

———. 2011. 'Peter Asaro: Military Robots and Just War Theory'. Federal Ministry of Defence (Austria). https://www.bmlv.org/pdf_pool/publikationen/20101105_et_ethical_and_legal_aspects_of_unmanned_systems_asaro.pdf.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. 'Synthesizing Robust Adversarial Examples'. arXiv:1707.07397. arXiv. https://doi.org/10.48550/arXiv.1707.07397.

Bellamy, Alex J. 2006. *Just Wars: From Cicero to Iraq*. Cambridge: Polity.

Blanchard, Alexander, and Mariarosaria Taddeo. 2022a. 'Jus in Bello Necessity, the Requirement of Minimal Force, and Autonomous Weapon Systems'. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4100042.

———. 2022b. 'Predictability, Distinction & Due Care in the Use of Lethal Autonomous Weapons Systems'. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4099394.

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

Boeing. 2022. 'Boeing: Autonomous Systems - Echo Voyager'. Boeing. 2022. https://www.boeing.com/defense/autonomous-systems/echo-voyager/index.page.

Boulanin, Vincent, Laura Bruun, and Netta Goussac. 2021. 'Autonomous Weapon Systems and International Humanitarian Law'. Stockholm: Stockholm International Peace Research Institute.

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.

Boyle, Joseph M. 1978. 'Praeter Intentionem in Aquinas'. *The Thomist: A Speculative Quarterly Review* 42 (4): 649–65.

———. 1980. 'Toward Understanding the Principle of Double Effect'. *Ethics* 90 (4): 527–38.

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

Choudhury, L, Aoun A, Badawy D., de Alburquerque L. A., Marjane J., and Wilkinson A. 2021. 'Letter the Panel of Experts on Libya Established Pursuant to Resolution 1973 (2011) Addressed to the President of the Security Council'. s/2021/229. United Nations Security Council.

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.

Demy, Timothy J. 2020. '"Something Old, Something New" – Reflections on AI and the Just War Tradition'. In *Artificial Intelligence and Global Security*, edited by Yvonne R. Masakowski, 53–62. Emerald Publishing Limited. https://doi.org/10.1108/978-1-78973-811-720201003.

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.

Docherty, Bonnie. 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.

Draper, Kai. 2016. *War and Individual Rights: The Foundations of Just War Theory*. Oxford: Oxford University Press.

Floridi, Luciano. 2018. 'Soft Ethics and the Governance of the Digital'. *Philosophy & Technology* 31 (1): 1–8.

Floridi, Luciano, and Jeff W. Sanders. 2004. 'On the Morality of Artificial Agents'. *Minds and Machines* 14 (3): 349–79.

Galliott, Jai. 2017. *Military Robots: Mapping the Moral Landscape*. http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781317096009.

Gould, Harry D. 2014. 'Rethinking Intention and Double Effect'. *The Future of Just War: New Critical Essays, Ed. Caron E. Gentry and Amy E. Eckert (Athens: University of Georgia Press, 2014)*.

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

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, Douglas. 2019. 'Why Deep-Learning AIs Are so Easy to Fool'. *Nature* 574 (7777): 163–66. https://doi.org/10.1038/d41586-019-03013-5.

Heyns, Christof. 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. 2020a. 'The Black Box, Unlocked | UNIDIR'. 2020. https://unidir.org/publication/black-box-unlocked.

———. 2020b. 'The Black Box, Unlocked: Predictability and Understandability in Military AI'. United Nations Institute for Disarmament Research. https://doi.org/10.37559/SecTec/20/AI1.

———. 2021. 'Known Unknowns: Data Issues and Military Autonomous Systems'. United Nations Insitute For Disarmament Research.

Homayounnejad, Maziar. 2018. 'Ensuring Fully Autonomous Weapons Systems Comply with the Rule of Distinction in Attack'. In *Drones and Other Unmanned Weapons Systems under International Law*, 123–57. Brill Nijhoff.

Hurka, Thomas. 2010. 'The Consequences of War'. In *Ethics and Humanity: Themes from the Philosophy of Jonathan Glover*, edited by N. Ann Davis, Richard Keshen, and Jeff McMahan. Oxford: Oxford University Press.

ICRC. 2021a. 'Customary IHL - Rule 1. The Principle of Distinction between Civilians and Combatants'. International Committee of the Red Cross. 2021. https://ihl-databases.icrc.org/customary-ihl/eng/docindex/v1_rul_rule1#Fn_D70F41D7_00008.

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

International Committee of the Red Cross, ICR. 2019. '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.

Krishnan, Armin. 2009. *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Burlington: Ashgate.

Lang, Anthony F. 2016. 'Just War as Political Theory: Intention, Cause and Authority'. *Political Theory* 44 (2): 289–303.

Lazar, Seth. 2009. 'Responsibility, Risk, and Killing in Self-Defense'. *Ethics* 119 (4): 699–728. https://doi.org/10.1086/605727.

———. 2017. 'Evaluating the Revisionist Critique of Just War Theory'. *Daedalus* 146 (1): 113–24.

Lee, Steven. 2004. 'Double Effect, Double Intention, and Asymmetric Warfare'. *Journal of Military Ethics* 3 (3): 233–51.

MacIntosh, Duncan. 2021. 'Fire and Forget: A Moral Defense of the Use of Autonomous Weapons in War and Peace'. In *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare*, edited by Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, 9–23. Oxford University Press. https://philarchive.org/rec/MACFAF-5.

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.

McIntyre, Alison. 2001. 'Doing Away with Double Effect'. *Ethics* 111 (2): 219–55. https://doi.org/10.1086/233472.

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

McMahan, Jeff. 1994a. 'Innocence, Self-Defense and Killing in War'. *Journal of Political Philosophy* 2 (3): 193–221.

———. 1994b. 'Revising the Doctrine of Double Effect'. *Journal of Applied Philosophy* 11 (2): 201–12.

Moore, Cristopher. 1990. 'Unpredictability and Undecidability in Dynamical Systems'. *Physical Review Letters* 64 (20): 2354–57. https://doi.org/10.1103/PhysRevLett.64.2354.

Musiolik, Thomas Heinrich, and Adrian David 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.

Otsuka, Michael. 1994. 'Killing the Innocent in Self-Defense'. *Philosophy & Public Affairs* 23 (1): 74–94.

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

Quinn, Warren S. 1989. 'Actions, Intentions, and Consequences: The Doctrine of Double Effect'. *Philosophy & Public Affairs*, 334–51.

Ramzy, Austin. 2022. 'What Is Known about the Iranian-Made Drones That Russia Is Using to Attack Ukraine.' *The New York Times*, 17 October 2022, sec. World. https://www.nytimes.com/2022/10/17/world/europe/russia-ukraine-iran-drones.html.

Rice, H. G. 1956. 'On Completely Recursively Enumerable Classes and Their Key Arrays'. *Journal of Symbolic Logic* 21 (3): 304–8. https://doi.org/10.2307/2269105.

Riesen, Erich. 2022. 'The Moral Case for the Development and Use of Autonomous Weapon Systems'. *Journal of Military Ethics* 21 (2): 132–50. https://doi.org/10.1080/15027570.2022.2124022.

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*.

Rudin, Cynthia, Caroline Wang, and Beau Coker. 2020. 'The Age of Secrecy and Unfairness in Recidivism Prediction'. *Harvard Data Science Review* 2 (1). https://doi.org/10.1162/99608f92.6ed64b30.

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.

Scholz, Jason, and Jai Galliott. 2021. 'The Humanitarian Imperative for Minimally-Just AI in Weapons'. In *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare*, edited by Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, 55–72. Oxford: Oxford University Press.

Simpson, Thomas W., and Vincent C. Müller. 2016. 'Just War and Robots' Killings'. *The Philosophical Quarterly* 66 (263): 302–22.

STM. 2020. 'STM - KARGU - Rotary Wing Attack Drone Loitering Munition System'. STM. 2020. https://www.stm.com.tr/kargu-autonomous-tactical-multi-rotor-attack-uav.

Swiatek, Mark S. 2012. 'Intending to Err: The Ethical Challenge of Lethal, Autonomous Systems'. *Ethics and Information Technology* 14 (4): 241–54. https://doi.org/10.1007/s10676-012-9302-1.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. 'Intriguing Properties of Neural Networks'. *ArXiv:1312.6199 [Cs]*, February. http://arxiv.org/abs/1312.6199.

Taddeo, Mariarosaria, and Alexander Blanchard. 2022a. 'Ascribing Moral Responsibility for the Actions of Autonomous Weapons Systems'. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4096934.

———. 2022b. 'Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems—a Moral Gambit'. *Philosophy & Technology* 35 (3): 78. https://doi.org/10.1007/s13347-022-00571-x.

———. 2022c. 'A Comparative Analysis of the Definitions of Autonomous Weapons Systems'. *Science and Engineering Ethics* 28 (5): 37. https://doi.org/10.1007/s11948-022-00392-3.

Taddeo, Mariarosaria, Tom McCutcheon, and Luciano Floridi. 2019a. '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.

———. 2019b. '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. 2021a. 'Ethical Principles for Artificial Intelligence in National Defence'. *Philosophy & Technology*, October. https://doi.org/10.1007/s13347-021-00482-3.

———. 2021b. 'Ethical Principles for Artificial Intelligence in National Defence'. *Philosophy & Technology*, October. https://doi.org/10.1007/s13347-021-00482-3.

Taddeo, Mariarosaria, Marta Ziosi, Andreas Tsamados, Luca Gilli, and Shalini Kurapati. 2022. 'Artificial Intelligence for National Security: The Predictability Problem'. London: Centre for Emerging Technology and Security.

Uesato, Jonathan, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. 2018. 'Adversarial Risk and the Dangers of Evaluating Against Weak Attacks'. *ArXiv:1802.05666 [Cs, Stat]*, February. http://arxiv.org/abs/1802.05666.

Umbrello, Steven, Phil Torres, and Angelo F. De Bellis. 2020. 'The Future of War: Could Lethal Autonomous Weapons Make Conflict More Ethical?' *AI & SOCIETY* 35 (1): 273–82. https://doi.org/10.1007/s00146-019-00879-x.

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.

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

Whetham, David. 2010. 'The Just War Tradition: A Pragmatic Compromise'. In *Ethics, Law, and Military Operations*, edited by David Whetham. Basingstoke: Palgrave Macmillan.

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

Wyatt, Austin. 2020. 'So Just What Is a Killer Robot?: Detailing the Ongoing Debate around Defining Lethal Autonomy'. Washington Headquarters Services. 8 June 2020. https://www.whs.mil/News/News-Display/Article/2210967/so-just-what-is-a-killer-robot-detailing-the-ongoing-debate-around-defining-let/.