Philosophy & Technology (2021) 34:1707–1729
https://doi.org/10.1007/s13347-021-00482-3

RESEARCH ARTICLE

Ethical Principles for Artificial Intelligence in National Defence

Mariarosaria Taddeo[1,2] · David McNeish[3] · Alexander Blanchard[2] · Elizabeth Edgar[3]

Received: 5 June 2021 / Accepted: 22 September 2021 / Published online: 13 October 2021

Abstract

Defence agencies across the globe identify artificial intelligence (AI) as a key technology to maintain an edge over adversaries. As a result, efforts to develop or acquire AI capabilities for defence are growing on a global scale. Unfortunately, they remain unmatched by efforts to define ethical frameworks to guide the use of AI in the defence domain. This article provides one such framework. It identifies five principles—justified and overridable uses, just and transparent systems and processes, human moral responsibility, meaningful human control and reliable AI systems—and related recommendations to foster ethically sound uses of AI for national defence purposes.

Keywords Artificial intelligence · Control · Defence · Digital ethics · Ethical principles · Fairness · Just war theory · Responsibility · Reliability

1 Introduction

Maintaining a technological advantage has always been pivotal to the success of national defence measures. It is even more so in mature information societies (Floridi, 2016a). This is why over the past two decades, there have been growing efforts to design, develop and deploy digital technologies for national defence. Artificial intelligence (AI), in particular, has shown to have great potential to aid national defence measures. Indeed, scholars, policy-makers and military experts observe that there is an ongoing global race for the development of AI for defence (Taddeo & Floridi, 2018). For example, the latest national defence and innovation

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

[1] Oxford Internet Institute, University of Oxford, 1, St Giles, Oxford OX1 3JS, UK

[2] Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB, UK

[3] Defence Science Technology Laboratory (Dstl), Salisbury, UK

strategies of several governments—UK,[1] USA,[2] Chinese,[3] Singapore,[4] Japanese[5] and Australian[6]—explicitly mention AI capabilities, which are already deployed to improve the security of critical national infrastructures, like transport, hospitals, energy and water supply. NATO, as well, has identified AI as a key technology to maintain superiority over adversaries in its 2020 report on the future of the alliance (NATO, 2020).

The applications of AI in national defence are virtually unlimited, ranging from support to logistics and transportation systems to target recognition, combat simulation, training and threat monitoring. There is a growing expectation among military planners that AI could enable speedier and more decisive defeat of the adversary. As with the use in other domains, the potential of AI is coupled with serious ethical problems, ranging from possible conflict escalation, the promotion of mass surveillance measures and the spreading of misinformation to breaches of individual rights and violation of dignity. If these problems are left undressed, the use of AI for defence purposes risks undermining the fundamental values of democratic societies and international stability (Taddeo, 2014b, 2019a, b).

This article offers guidance to address these problems by identifying ethical principles to inform the design, development and use of AI for defence purposes. These principles should not be taken as an alternative to national and international laws; rather, they offer guidance to the use of AI in the defence domain in ways that are coherent with existing regulations. In this sense, the proposed principles indicate what ought to be done or not to be done.

> "*over and above* the existing regulation, not against it, or despite its scope, or to change it, or to by-pass it (e.g. in terms of self-regulation)" (Floridi, 2018, p. 4).

In offering these principles, the paper fills an important gap in the relevant academic and policy literature; while numerous ethical principles and frameworks have been published which focus on civilian applications of AI (Jobin et al., 2019), very few so far address directly the problems inherent to the use of AI in the defence domain. In the rest of this article, Sect. 2 describes the methodology used for our analysis. Section 3 offers an analysis of the ethical problems linked to current uses of this technology in the defence. Section 4 focuses on the ethical principles provided by the US Defense Innovation Board (DIB). Thus far, these are the only domain-specific principles published by a defence institution. The analysis of these principles shows some key limitations of the DIB approach and paves the way to Sect. 5,

[1] https://www.gov.uk/government/publications/future-force-concept-jcn-117

[2] https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

[3] (Roberts et al., 2020).

[4] https://www.csa.gov.sg/~/media/csa/documents/publications/singaporecybersecuritystrategy.pdf

[5] https://www.nisc.go.jp/eng/pdf/cs-senryaku2018-en.pdf

[6] https://www.business.gov.au/news/budget-2019-20

which introduces five new ethical principles to guide the use of AI for national defence. Section 6 concludes the article.

2 Methodology

The first steps to identifying viable ethical principles to guide the use of AI for national defence are the definition of AI and the identification of the ethical problems that its use may pose. For the purposes of our analysis, we can abstract from specific technical aspects of AI systems (we can disregard, for example, whether the system under analysis is a statistical or a subsymbolic one) and adopt a high level of abstraction (LoA) (Floridi, 2008). Thus, we consider AI as.

> "a growing resource of interactive, autonomous, and self-learning agency, which can be used to perform tasks that would otherwise require human intelligence to be executed successfully" (Floridi & Cowls, 2019).

The choice of the method to identify the ethical problems posed by the use of AI for national defence is not a trivial one. For example, one may think of developing a complete taxonomy of the ethical issues posed by existing uses of AI in the defence domain, but this is unfeasible and of little value: the taxonomy would be quickly outdated by the rapid developments in AI and its application to new uses. At the same time, analyses that disregard the specific domain and purpose of deployment risk defining ethical principles which are too generic to provide any concrete guidance.

Hence, the choice of LoA becomes crucial to develop a correct analysis of the ethical problems and define principles able to provide actionable guidance. Given the goal of this article, we chose a gradient of analysis (GoA) that combines two LoAs: $\text{LoA}_{\text{purpose}}$ and $\text{LoA}_{\text{ethics}}$. The observables of $\text{LoA}_{\text{purpose}}$ are the purposes of deployment of AI. The observables of $\text{LoA}_{\text{ethics}}$ are, for any given purpose, the aspects of the design, development and deployment of AI that may lead to un/ethical consequences.

The decision to focus on purposes of use rather than on the function of the technology requires clarification. It rests on two reasons. First is the dual-use nature of AI—as with any digital technology, AI is *malleable* and its original function can be easily repurposed. Hence, un/ethical implications of its uses are not necessarily defined by their design function as much as they are determined by the purpose with which these technologies are deployed. Second, within the defence domain, these purposes can be clearly identified and are likely to shape both current and future uses of AI, and thus, ethical principles that focus on purposes of use, rather than on the specific function, of a given technology have a better-defined scope and their guidance is more likely to stand the test of time.

The LoAs embraced for this analysis have a medium granularity. Thus, they identify problems (and inform the definition of principles in Sect. 5) that are specific to the domain but do not distinguish among specific contexts (e.g. naval or aviation) of AI deployment within the defence domain and hence disregard the variation of ethical challenges that may occur between contexts. Consider, for example, the

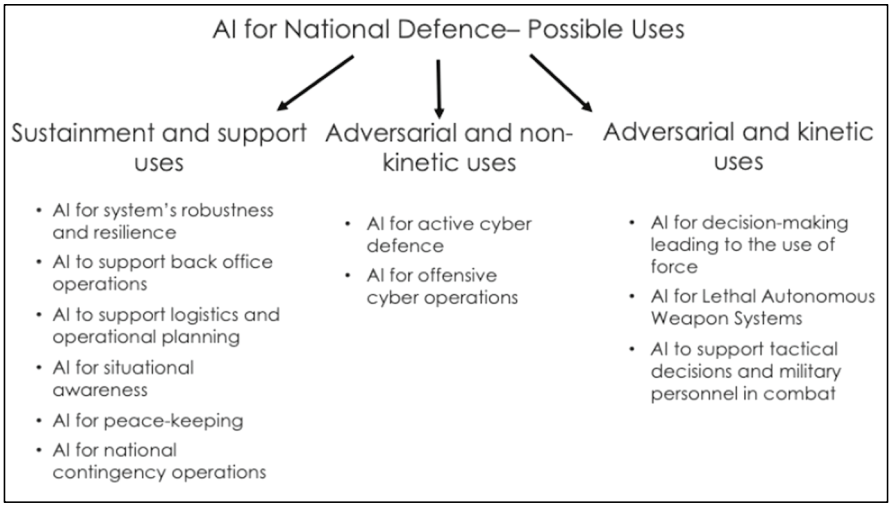

Fig. 1 The three purposes of use of AI for national defence

different ethical problems and related solutions for using AI in submarines or aviation operations.

Purposes of use of AI in defence span over three core categories of action by defence institutions (Fig. 1): sustainment and support, adversarial and non-kinetic, adversarial and kinetic. We shall delve into the ethical implications of each of these in Sect. 3, but let us describe them briefly here. Sustainment and support uses of AI refer to all cases in which AI is deployed to support 'back-office' functions, as well as logistical distribution of resources. This category also includes uses of AI to improve the security of infrastructure and communication systems underpinning national defence. Adversarial and non-kinetic uses of AI range from uses of AI to counter cyber-attacks to active cyber defence, and offensive cyber operations with non-kinetic aims. Adversarial and kinetic uses refer to the integration of AI systems in combat operations, and these range from the use of AI systems to aid the identification of targets to lethal autonomous weapon systems (LAWS).

The ethical principles for the use of AI in the defence domain that we provide in this article refer to sustainment and support uses and to adversarial and non-kinetic uses of AI. The ethical analysis of the use of AI for adversarial and kinetic purposes will be addressed in the second stage of our research.

3 Ethical Challenges of AI for Defence Purposes

Figure 2 shows the minimum requirements, for each purpose of use of AI in defence that AI systems should meet to be ethically sound.

The three purposes of use of AI in the defence domain are more ethically problematic as one moves from sustainment and support uses to adversarial and kinetic uses. This is because alongside the ethical problems related to the use of AI (e.g.

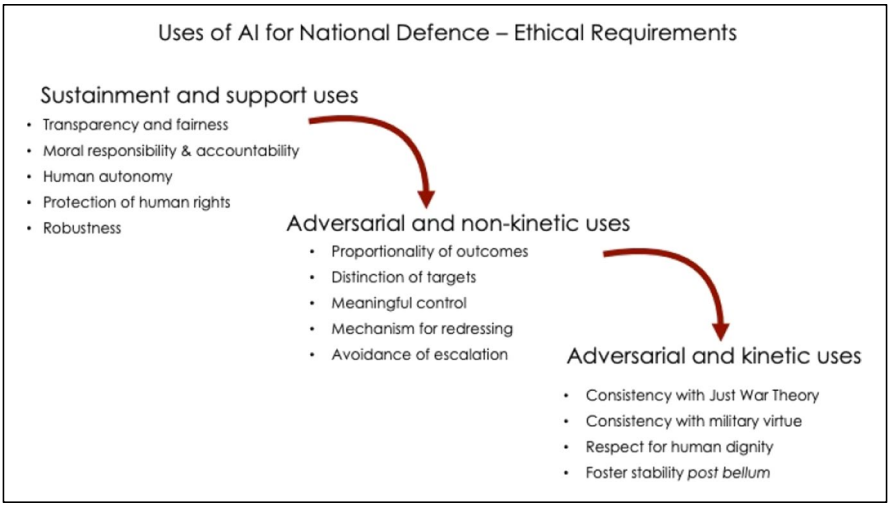

Fig. 2 A map of the ethical requirements linked to the specific purpose of the use of AI in defence

transparency and fairness), one also needs to consider the ethical problems related to adversarial, whether non-kinetic or kinetic, uses of this technology and its disruptive and destructive impact.

As shown in Fig. 2, each category of use has its own specific ethical requirements but also inherits the ones from the categories on its left. For example, to be ethically sound, adversarial and non-kinetic uses of AI need to ensure some forms of meaningful control and measures to avoid escalation, while also respecting transparency and human autonomy, which appear in the sustainment and support category. Some AI systems have dual capability and can be used both defensively and offensively. Independent of the capability in which they are used, the systems still need to meet the requirements specified in Fig. 2. For example, whether in an offensive or defensive operation, uses of the systems need to be accountable, proportionate and coherent with the principles of the just war theory.

Let us now consider in more details some of the key ethical problems of each purpose of use.

3.1 Sustainment and Support Uses of AI

Defence organisations already employ AI systems for different non-adversarial aspects of operations (US Army, 2017). Uses vary from applications in cybersecurity, where AI plays an ever-growing role to ensure systems robustness and resilience, to AI-based drones capturing video reconnaissance (Lysaght et al., 1988; Fraga-Lamas et al., 2016; Schubert et al., 2018).

For nations with technically advanced militaries, AI systems are likely to be fully integrated into national defence capabilities to support back-office, logistics and security tasks. For example, research estimates that the number of intelligent sensors in a military setting could reach one million per square kilometre similar to

the supported connection density of the 5G network (Kott et al., 2017; International Telecommunications Union 2017). This has been described as *the Internet of battle things* (Kott et al., 2017). In these cases, AI will play a key role to ensure the robustness and resilience of the networks as well as to elaborate data and extract, filter, collate, cross-link and communicate relevant information. All these uses pose serious ethical problems.

First, consider the use of AI to enhance system robustness. This refers to AI for software testing, which is a new area of research and development. It is defined as an.

> "emerging field aimed at the development of AI systems to test software, methods to test AI systems, and ultimately designing software that is capable of self-testing and self-healing".[7]

In this sense, AI can take software testing to a new level, making systems more robust (King et al., 2019). However, delegating testing to AI could lead to a complete deskilling of defence personnel deployed for verification and validation of systems and networks and a subsequent lack of control of this technology (Yang et al. 2018; Taddeo, 2019a, b).

Next, let us focus on system resilience. AI is increasingly deployed for threat and anomaly detection (TAD). TAD can make use of existing security data to train for pattern recognition. As stressed by Taddeo et al. (2019) in some cases, threat scanners have access to files, emails, mobile and endpoint devices, or even traffic data on a network. Monitoring extends to users as well. AI can be used to authenticate users by monitoring behaviour and generating biometric profiles, like, for example, the unique way in which a user moves her mouse around (BehavioSec: Continuous Authentication Through Behavioral Biometrics 2019). In this case, the risk is clear. This use of AI puts users' privacy under a sharp devaluative pressure, exposing users to extra risks should data confidentiality be breached and creating a mass-surveillance effect (Taddeo, 2013, 2014b).

AI can extract information to support logistics and decision-making, but also for foresight analyses, internal governance and policy. These are perhaps some of the uses of AI with the greater potential to improve defence operations, as they will facilitate timely and effective management of both human and physical resources, improve risk assessment and support decision-making processes. For example, a report by KPMG[8] stresses that a defence agency could have only a few minutes to decide whether a missile launch represents a threat, share the findings with allies and decide how to respond. AI would be of great help in this scenario, for it could integrate real-time data from satellites and sensors and elaborate key information that may facilitate and improve human decision-making process by mitigating uncertainties due to the fog of war and possible human biases. The challenge is that these uses of AI must ensure that the systems would not perpetrate a biased decision and

[7] www.aitesting.org

[8] https://assets.kpmg/content/dam/kpmg/xx/pdf/2018/04/next-major-defense-challenge.pdf

unduly discriminate, while also offering a means to maintain accountability, control and transparency.

3.2 Adversarial and Non-kinetic Uses of AI

As cyber threats escalate, so does the need for defence strategies required to meet them. The UK and the USA have employed *active* cyber defence strategies that enable computer experts to neutralise or distract viruses with decoy targets, and to break back into a hacker's system to delete data or to destroy it completely. In February 2020, the UK also established the National Cyber Force, as a joint initiative between the Ministry of Defence and GCHQ, which is tasked to target hostile foreign actors. On an international scale, NATO can now rely on sovereign cyber effects in response to cyber-attacks, as agreed at the Brussels Summit.[9] This may enable the alliance to punish (attributed) attacks and deter attackers from striking again in the future (Taddeo, 2019a).

AI will revolutionise these activities. Attacks and responses will become faster, more precise and more disruptive. It will also expand the targeting ability of attackers, enabling them to use more complex and richer data. Enhancing current methods of attack is an obvious extension of existing technology; however, using AI within malware can change the nature and delivery of an attack. Autonomous and semi-autonomous cybersecurity systems endowed with a "playbook" of pre-determined responses to an activity, constraining the agent to known actions, are already available on the market (DarkLight Offers First of Its Kind Artificial Intelligence to Enhance Cybersecurity Defenses, 2017). Autonomous systems able to learn adversarial behaviour and generate decoys and honeypots (Acalvio Autonomous Deception 2019) are also being commercialised. Additionally, AI-enabled cyber weapons have already been prototyped including autonomous malware, corrupting medical imagery and attacking autonomous vehicles (Mirsky et al., 2019; Zhuge et al., 2007). For example, IBM created a prototype autonomous malware, DeepLocker, that uses a neural network to select its targets and disguise itself until it reaches its destination (DeepLocker: How AI Can Power a Stealthy New Breed of Malware, 2018).

As states use increasingly aggressive AI-driven strategies, opponents may respond more fiercely (Taddeo and Floridi 2018). This may expand into an intensification of cyber-attacks and responses, which, in turn, may pose serious risks of escalation and lead to kinetic consequences (Taddeo 2017). To avoid the escalation, it is vital that uses of AI respect key principles of the just war theory which underpins international regulations (Taddeo, 2012a, b, 2014a), such as the United Nations Charter,[10] The Hague and Geneva Conventions[11] and international humanitarian law,[12] and sets the parameters for both ethical and political debates on waging

[9] https://www.nato.int/docu/review/articles/2019/02/12/natos-role-in-cyberspace/index.html
[10] https://www.un.org/en/sections/un-charter/un-charter-full-text/
[11] https://www.loc.gov/rr/frd/Military_Law/pdf/ASubjScd-27-1_1975.pdf
[12] https://www.icrc.org/en/doc/resources/documents/misc/57jm93.htm

conflicts. It is crucial that the deployment of AI for aggressive and non-kinetic purposes respects the principles of proportionality of responses, discriminates between legitimate and illegitimate targets, ensures some form of redress when mistakes are made (Taddeo, 2012a, b, 2014a) and maintains responsibility and control within the chain of command. Ultimately, ethical analyses of the adversarial and non-kinetic use of AI should contribute to understanding how to apply the just war theory in cyberspace and be used to shape the debate on the regulation of state behaviour in this domain (Taddeo & Floridi, 2018).

3.3 *Adversarial and Kinetic Uses of AI*[13]

The use of AI for aggressive and kinetic purposes varies, ranging from automating various functions of a weapon system to systems that follow the pre-programmed instructions of a human and to full autonomy, where the weapons system identifies, selects and engages targets without any human input. Consider, for example, STARTLE,[14] a system developed for the Royal Navy to support human decision-making. It is endowed with situational awareness software that monitors and assesses potential threats using a combination of AI techniques. Similarly, the Advanced Targeting and Lethality Automated System (ATLAS)[15] developed for the US Army supports humans in identifying threats and prioritises potential targets. Ethical problems vary with the degree of autonomy of weapon systems, the level of force that they can deploy and the nature of the possible targets, whether material or humans.

While many countries have expressed their commitment not to develop or use fully autonomous weapon systems, it is still important to consider and address the ethical problems they pose in order to establish boundaries for the development and use of weapons which incorporate AI but are not fully autonomous in their operation or may not target human agents.

A key challenge is to ensure that adversarial and kinetic uses of AI will be able to respect the tenets of the just war theory, for example necessity, proportionality and discrimination. So, for example, AI systems must be able to distinguish between combatants and non-combatants carrying a weapon or recognising the generally accepted signs of surrender that operate in armed conflict. This may be problematic, because AI, at least in its current state of development, is insufficiently able to analyse context; in some situations, its capacity to recognise who is and who is not a legitimate target could be significantly worse than that of humans (Sharkey, 2010, 2012a, b; Tamburrini, 2016).

The responsibility gap is another key ethical challenge. As mentioned in Sect. 3, while a responsibility gap is problematic in all the three categories of use of AI, it

[13] As mentioned in Sect. 2, in this article, we do not focus on adversarial and kinetic uses of AI, as this is the focus of a second stage of our work. Nonetheless, in this section, we offer an overview of the ethical problems related to this use, with the goal of providing a comprehensive overview of the ethical problems of using AI in defence.

[14] https://www.roke.co.uk/products/startle

[15] https://breakingdefense.com/2019/03/atlas-killer-robot-no-virtual-crewman-yes/

is particularly worrying when considering the adversarial and kinetic case, given the high stakes involved (Sparrow, 2007). This gap becomes even more pressing when coupled with the respect of the opponent and of her dignity. Treating opponents with respect in warfare is an important way of maintaining warfare's morality (Nagel, 1972), and the interpersonal relation with the opponent is considered to be a key to this end. Insofar as the use of autonomous LAWS would sever this relation, they undermine the dignity of those whom they target and lead to a form of morally problematic killing (Asaro, 2012; Docherty, 2014; Ekelhof, 2019; Johnson & Axinn, 2013; O'Connell, 2014; Sharkey, 2019; Sparrow, 2016).

Finally, questions arise with respect to the impact of LAWS on international stability. On the one side, LAWS may reduce the time span of the hostilities in which states may engage and thus contribute to fostering stability. They could also be an effective deterrent against possible opponents. On the other side, LAWS may lead to unjust war and hamper international instability. Some argue that this is because the use of LAWS may lower the barriers to warfare (Brunstetter & Braun, 2013; Enemark, 2011), possibly increasing the number of wars. For instance, it may be the case that the widespread use of LAWS would allow decision-makers to wage wars without the need to overcome the potential objections of military personnel (McMahan, 2013). In the same vein, asymmetric warfare that would result from one side using LAWS may lead to the weaker side, resorting to insurgency and terrorist tactics more often (Sharkey, 2012a, b). Because terrorism is considered to be a form of unjust warfare (or, worse, an act of indiscriminate murder), deploying LAWS may lead to a greater incidence of immoral violence.

4 Ethical Guidelines for the Use of AI

Over the past few years, several frameworks for the ethical design, development and use of AI have been proposed (Floridi & Cowls, 2019). For example, Jobin et al. (2019) identified 84 ethical frameworks for AI in their review. Ethical guidelines can vary in a number of dimensions, e.g. by the agency putting them forward (from governments to non-governmental organisations), by the scope of their application (e.g. from guidelines for private sector, e.g. social media, to guidelines for all entities developing and using AI, e.g. the European Guidelines for Trustworthy Artificial Intelligence) and by the applications they are seeking to govern (e.g. from national defence applications to applications in the wider public sector).

Despite the wide scope covered by existing frameworks and despite the long tradition of military ethics and the just war theory, uses of AI in the defence domain (especially the non-kinetic) have received very little attention in both the policy and academic literature. Thus far, the principles defined by the US DIB (2020a) for the use of AI in defence are the only exception to this lack of focus. In this section, we analyse the DIB principles to identify both their points of strength and limitations. The goal is to extract valuable lessons to learn before moving to describe the principles that we propose in this article.

The DIB identifies five principles: responsible, equitable, traceable, reliable and governable AI; they are meant to be applied to both kinetic and non-kinetic

uses of AI, whether adversarial or not. The following subsections analyse each principle, in turn, focusing on both the principles described in DIB (2020a) and the wider supporting report (DIB, 2020b), where the DIB describes the rational for each principle and specific recommendations for their implementation.

4.1 Responsible

The DIB principle states that.

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

This principle is uncontroversial and coherent with other ethical frameworks (Department for Digital, Culture, Media & Sport 2018, 5; Gavaghan et al., 2019, 41; Japanese Society for Artificial Intelligence [JSAI] 2017, p. 3).

In the supporting document, the recommendation on how to implement this principle proposes a three-level system of responsibilities, with the first level addressing humans who control.

> "the design, requirements definition, development, acquisition, testing, evaluation, and training for any DoD system, including AI ones" (DIB, 2020b, p. 27).

The second level addresses the use of AI in the conduct of hostilities (whether kinetic or not); in this case, responsibilities are ascribed according to the command and control structure, insofar as commanders and operators have "appropriate information on a system's behaviour, relevant training, and intelligence and situational awareness" (p. 28). The third level of responsibility refers to redressing mechanisms for actions after hostilities have ended. This level addresses both the Department of Defense (DoD) and private sector procuring AI technology. The DIB supporting documents specify that human responsibility rests on 'human appropriate judgment'.

This approach is correct only in part. There are two main limitations to it. On the one side, the definition of 'appropriate' judgement remains vague and, therefore, problematic especially when considering the problems posed by the lack of transparency and predictability of some AI systems. On the other side, the attribution of responsibility according to the three-level system risks dumping responsibilities on the first level, insofar as unintended consequences of AI systems can, in majority of the cases, be linked back to design and development issues. This may have a detrimental effect on the way actors involved in command and control may perceive their responsibilities with respect to the use of AI.

4.2 Equitable

The DIB principle prescribes that.

> "The DoD should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons" (p. 8).

This principle focuses on issues related to fairness and justice; however, it avoids referring to the two concepts directly. In the supporting document, the reason given for not using the term 'fairness' in the principle is the following:

> "this principle stems from the DoD mantra that fights should not be fair, as DoD aims to create the conditions to maintain an unfair advantage over any potential adversaries, thereby increasing the likelihood of deterring conflict from the outset" (DIB, 2020b, p. 31).

The document goes on to say that the.

> "DoD should have AI systems that are appropriately biased to target certain adversarial combatants more successfully and minimize any pernicious impact on civilians, non-combatants, or other individuals who should not be targeted" (Defence Innovation Board (DIB) (2020b, p. 33).

This departure, then, is motivated by the perceived unique nature of defence AI applications. When considering fairness with respect to AI, the DIB principles centre only on the unfair impact of the use of AI on the DoD personnel, disregarding the problems that the lack of fairness may pose when deploying AI on the cyber and kinetic battleground. This is misleading, as it may suggest that the need to seek advantage over the adversary may justify unfair, or indeed unjust, practices. This is not the case, as we distinguish between just and unjust conduct in defence and punish the latter.

There are differences between the ways in which the principle of justice is applied in civilian and non-belligerent contexts and in hostile activities. The just war theory and international humanitarian law define the terms of this principle and how to respect it when conducting hostilities. These terms differ, at times radically, from the ones referring to civilian uses, but still define the space of just conduct—and hence fairness—in defence. Ethical guidelines for AI in defence need to define principles for just uses of AI which are relevant within this domain and coherent with the principles provided by the just war theory (Taddeo, 2014a).

4.3 Traceability

This principle addresses *indirectly* the ethical problems posed by the lack of transparency of AI. It states that.

> "DoD's AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation" (DIB, 2020a, p. 8).

Notably, the focus of the principle is not on the transparency of the technology but on the skills of the DoD personnel and their level of understanding of AI systems, insofar as these facilitate the traceability of the processes and decisions of AI systems at both development and deployment stages. As specified in the supporting document, traceability at development stage refers to the collection and sharing with appropriate stakeholders of "design methodology, relevant design documents, and data sources" (p. 34), whereas at deployment stage, traceability includes forms of monitoring, auditing and transparency of processes. As specified in the DIB supporting document:

> "Some systems may require not just reviews of user access, but also records of use and for what purpose. This requirement can mitigate harms related to off-label use of an AI system, as well as reinforce the principle of responsibility. In short, DoD will need to rethink how it traces its AI systems, who has access to particular datasets and models, and whether those individuals are reusing them for other application areas" (p. 35).

While analysis provided in the supporting document links correctly the transparency of processes to responsible uses of AI, it overlooks the relation between transparency of AI and human responsibilities. Indeed, the event of mistakes, malfunctioning or unintended consequences following from the use of AI, traceability of processes and decisions may compensate for the lack of transparency of this technology. However, albeit useful, the approach adopted with this principle offers a remedy, not a solution to the challenges posed by the lack of transparency of AI. Traceability without transparency is very limited. While it may foster responsible uses of AI, it does not shed much light on the responsibilities for mistakes and failures of this technology, nor does it offer an opportunity to identify promptly the sources of mistakes and unwanted outcomes of AI systems. The DIB documents do not stress so and do not propose any suggestions to overcome the lack of transparency of AI system.

4.4 Reliable and Governable

The principle focusing on reliability of AI states that.

> "DoD AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use" (p. 8).

This principle resonates with one of the principles provided by Organisation for Economic Co-operation and Development (OECD), which stresses that.

> "AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed".[16]

[16] https://www.oecd.org/going-digital/ai/principles/

In the supporting document, the DIB (2020a, b) stresses the need for reliable AI (rather than trustworthy), whose "safety, security, and robustness […] should be tested and assured" (DIB, 2020a, p. 8). This principle is specifically oriented at fostering verification and validation as well as to improve AI robustness. We believe that this is a crucial requirement for the use of AI in defence and the one that is important to mention explicitly to reiterate the need to monitor AI systems, especially when these are deployed within a defence organisation (more on this in Sect. 5).

At the same time, the DIB supporting documents highlight the importance of human agents being able to disengage or deactivate systems that demonstrate unintended escalatory behaviour. The supporting document emphasises the need for human control given the unpredictable behaviour of some AI systems, especially those operating in complex and dynamic environments (DIB, 2020b, p. 39).[17]

Control is not mentioned explicitly in the principles, but it is central to the principle focusing on governable AI, which prescribes that.

> "DoD AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior".

While pointing at the correct direction, insofar as it specifies the need to maintain AI under some forms of control, the principle remains vague with respect to what the desirable forms of control should be, and how this should be exerted, and what the minimum level of ethically acceptable control is.

These are important aspects to consider, especially as defence institutions increasingly deploy AI in hybrid teams, including human and artificial agents. In this scenario, a *governable* AI offers too generic a guidance to identify, for example, ethically sound forms of control over AI systems or how to attribute responsibilities for failures with respect to misuses or overuses of this technology.

In this respect, a notable omission in the DIB principles is the lack of focus on human autonomy. While this is acceptable insofar as autonomy may be considered a principle attaining to personal sphere and the ability of individuals to pursue their own choices, this is also a missed opportunity. Autonomy protects individuals' ability to dissent from AI-based decisions. In this sense, as AI is increasingly embedded in the decision-making processes of defence organisations, it is important to protect the ability of human agents to contest and override AI decisions, when these should be considered mistaken or inappropriate. In this sense, the principle of autonomy enables stronger forms of control over the use of AI.

[17] It should be noted that the High-Level Expert Group's principles also include provisions for human control, but given its focus on trustworthy AI, these are more flexible. For example, it allows that less human oversight may be exercised so long as more extensive testing and stricter governance is in place.

5 Five Ethical Principles for Sustainment and Support and Adversarial and Non-kinetic Uses of AI

In this section, we offer five ethical principles specifically designed to address the ethical problems posed by the deployment of AI in the defence domain. The principles specified in this article refer to both sustainment and support uses and adversarial and non-kinetic uses of AI. They should be regarded as the first building block of a more comprehensive ethical framework addressing also the adversarial and kinetic uses of AI, which will be the focus of the second, forthcoming, part of this project.

In order to be ethically sound, sustainment and support and adversarial and non-kinetic uses of AI for national defence purposes should respect the following ethical principles:

i. Justified and overridable uses
ii. Just and transparent systems and processes
iii. Human moral responsibility
iv. Meaningful human control
v. Reliable AI systems

5.1 Justified and Overridable Uses

The (non) adoption of AI needs to be justified to ensure that AI solutions are not being underused, thus creating opportunity costs, or overused and misused, thus creating risks. Similarly, the decision to (or not to) resort to AI should always be overridable, should it become clear that it leads to unwanted consequences.[18]

Even when designed and deployed according to ethical principles, AI remains an ethically challenging technology. Its use may lead to great advantages for national defence. Yet, AI is not a silver bullet. This is a lesson that should be learned from the ethical governance of AI for social good. As Floridi and colleagues (2020, p. 1773) stress:

> "it is important to acknowledge at the outset that there are myriad circumstances in which AI will not be the most effective way to address a particular social problem. This could be due to the existence of alternative approaches that are more efficacious or because of the unacceptable risks that the deployment of AI would introduce".

At the same time, AI can also encroach upon human rights and international humanitarian law or pose risks to international stability (the reader will recall the risks of the snowball effect linked to the adversarial and non-kinetic use of AI). This

[18] There are cases where the just war theory does allow unwanted consequences. Under the Doctrine of Double Effect (DDE), the just war theory permits unintended but foreseeable harms to non-combatants. However, the unwanted consequences entailed by DDE are so strictly ringfenced by the principles of proportionality and necessity, as well as the combatant obligations of due care, that the set of unwanted consequences actually permitted by the just war theory is exceptionally limited.

is why the decision to (or not to) delegate tasks to AI systems should follow a careful analysis of the ethical risks and benefits in any given context of deployment to justify it.

This principle yields different recommendations when considering sustainment and support and adversarial and non-kinetic uses. In the first case, the principle calls for an assessment of the ethical risks against the expected benefits following from the deployment of AI systems, for example weighting the benefits of using an AI system that may speed up a decision-making process or optimise logistic and distribution of resources against the likelihood that it may have a negative impact on jobs and human expertise, or considering the impact on human autonomy when AI is integrated in human teams (human–machine teaming).

When deciding on deploying AI for adversarial and non-kinetic purposes, for example for offensive cyber operations, it is essential to ensure that AI systems will respect the principles of necessity, humanity, distinction and proportionality (The UK and International Humanitarian Law 2018, n.d.). This may prove to be a complex task, as the principles of international humanitarian law, and the underpinning principles of the just war theory, are geared toward kinetic forms of conflicts, and therefore, their implementation to the case of non-kinetic warfare may be problematic. Consider, for example, proportionality and the problems of assessing the expected damage to intangible entities (e.g. data or services) against the concrete military aim to be achieved (Taddeo, 2012a, b, 2014a). Satisfying this principle will require extending the scope of the fundamental tenets of the just war theory from kinetic to non-kinetic operation, a complex but necessary, and not impossible, task.

Given the learning capability of AI and the potential lack of predictability of its outcome, even when uses of AI are justified, a constant monitoring of the ethical soundness of the solutions that they provide should be in place. Similarly, procedures to override the decision to resort to AI in a timely and effective way should be established every time an AI system is deployed.

5.2 Just and Transparent Systems and Processes

AI systems should not perpetrate any undue discrimination, nor should they lead to any breach of the principles of the just war theory. To this end, AI defence institutions should ensure that the deployed AI systems, and the processes in which they are embedded, remain transparent (and explicable) to facilitate the identification of the origin of any breach of the principles of the just war theory, of unintended and mistaken outcomes, the attribution of responsibilities, and guarantee the possibility of scrutinising and challenging processes and outcomes to ensure that they remain ethically sound.

Three aspects are crucial:

- Establish processes for ethical auditing
- Ensure that developed and procured AI systems are deployed in ways that respect the principles of the just war theory

- Maintain traceability for the design, development or procurement and deployment of AI systems

Ethical auditing should involve the entire decision-making process, and so, it should focus on both human and artificial agents, to ensure that both agents respect the relevant ethical principles (Mökander & Floridi, 2021).

Transparency of AI systems and processes enables access to the relevant information. The former requires explainability, while the latter traceability. Transparency of AI follows from the effort of designing and developing explainable technologies. Thus, it is crucial that in-house and procured AI systems are designed and developed with explainability in mind. Defence agencies should consider participating actively in the 'design-develop-deploy' cycle of the AI technologies that they procure and contribute to the development phase by setting standards and offering a trusted space where these technologies could be beta-tested. To facilitate this process, procurement policies should account for an ethical scrutiny of the third parties involved. While for national interest and security it is likely that scrutiny in this area may not be public, it is important that it is conducted by independent bodies or committees, which should be enabled and supported to develop an objective, in-depth assessment and should be accountable to the public for their assessment.

AI systems are often designed and developed in a distributed way, and models, data, training and implementation may be managed by different actors. At the same time, AI learns by experience: past deployments can impact future outcomes. This is why transparency requires traceability of sourcing and practices, to ensure that the chain of events leading to possible unwanted outcomes is not lost in the distributed and dynamic nature of design, development and deployment of AI.

5.3 Human Moral Responsibility

Humans remain the only agents morally responsible for the outcomes of AI systems deployed for defence purposes. While AI systems can be considered moral agents, insofar as they perform actions that have a moral value (Floridi & Sanders, 2004), they cannot be held morally responsible for those actions.

However, ascribing responsibilities to humans for the actions of AI systems has proved to be problematic, due to the distributed and interconnected ways in which AI is developed and the lack of transparency and predictability of its outcomes. Two approaches can be followed to enable fair processes to ascribe responsibilities:

- Following the chain of command, control and communication
- Faultless, back-propagation approach

They can be described more simply as a 'linear' approach and a 'radial' approach, respectively. These two approaches are complementary and serve the twin purposes of addressing unwanted consequences, misuses and overuses of AI and to foster a self-improving dynamic in the network of agents involved in the design, development and deployment of AI for defence.

According to the linear approach, responsibility is attributed following the chain of command, control and communication. In this case, decision-makers are held responsible for the unwanted consequences of AI, whether these result from failures of AI systems, unpredictability of outcomes or bad decisions. In order to ascribe responsibility fairly, it is essential that the decision-makers have adequate information and *understanding* of the way the specific AI system works in the given context, of its robustness, of the risks that it may deliver unpredicted (and unwanted) outcomes, of the required level of meaningful control and of the dangers that may follow if the AI systems fails to behave according to expectations. The linear approach entails a certain epistemic threshold. This means that the use of AI must be coupled with proper training of the personnel, both those who decide to deploy AI systems and those who use it, so that they understand the ways in which AI systems work, the risks and benefits linked to the systems and the ethical and legal implications of the decision to deploy AI. This approach rests on the idea that informed decision-makers choosing to use AI do so while being aware of the risks that this may imply and take responsibility for it.

The radial approach is useful to address unwanted outcomes of AI systems that do not stem from the intentions of human agents or follow from actions that are morally neutral per se. This approach addresses unethical consequences that spur from the convergence of different, independent, morally neutral factors. In the relevant literature, this has been defined as *faultless responsibility* (Floridi, 2016b). It refers to contexts in which, while it is possible to identify the causal chain of agents and actions that led to a morally good/bad outcome, it is not possible to attribute intent to perform morally good/bad actions to any of those agents individually and, therefore, all the agents are held morally responsible for that outcome insofar as they are part of the network which determined it.

This is not an entirely new approach, as it is akin to the legal concept of strict liability. According to strict liability, legal responsibility for unwanted outcomes is attributed to one or more agents for the damage caused by their actions or omissions, irrespective of the intentionality of the action and feasibility of control. When considering human–machine teaming—the integration of AI systems in defence infrastructures, decision-making processes and operations—what one needs to show to attribute moral responsibility according to the radial approach is that.

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

All the agents of the network are then held maximally responsible for the outcome of the network. As Floridi (2016b) stresses, this approach does not aim at distributing reward and punishment for the actions of a system, rather it aims at establishing a feedback mechanism that incentivises all the agents in the network to improve its outcomes—if all the agents are morally responsible, they may become more cautious and careful and this may reduce the risk of unwanted outcomes. This becomes quite effective when, for example, the moral responsibility is linked to the reputation of the agents.

5.4 Meaningful Human Control

It follows from the previous principle that the deployment of AI should always envisage meaningful forms of human control. Meaningful forms of human control must be in place to limit the risks that the outcome of AI systems will not meet the original intent, to identify promptly mistakes and unintended consequences, as well as to ensure timely intervention on, or deactivation of, the systems, should this be necessary.

The concept of meaningful control has been discussed widely in the relevant literature on LAWS, and indeed when considering these systems, control is a key element to consider. However, meaningful control is necessary also when considering uses of AI that may not lead to the use of force. This is because.

> "military systems must be able to function safely and effectively under a wide range of highly dynamic environments and use cases that are hard to predict or anticipate during the design phase. They must also be resilient to failure and to complex, uncertain and unpredictable events and situations where the dynamics of the military domain necessitate complex judgements regarding acceptable actions based on rules of engagement, international law and judgements over legality, proportionality and risk. Because of this the maintenance of Human Control through a combination of specification, design, training, operating procedures, and assurance processes is seen as critical in many, if not all military systems" (Boardman & Butcher, 2019, p. 2).

Meaningful human control of AI is characterised as dynamic, multidimensional and situation-dependent, and it can be exercised focusing on different aspects of the human–machine team. For example, the Stockholm International Peace Research Institute and the International Committee of the Red Cross identify three main aspects of human control of weapon systems: the weapon system's parameters of use, the environment and human–machine interaction (Boulanin et al., 2020). More aspects can also be considered. For example, Boardman and Butcher (2019) suggest that control should not just be meaningful but *appropriate*, insofar as it should be exercised in such a way to ensure that the human involvement in the decision-making process remains significant without impairing system performance.

While meaningful control can be dynamic, multidimensional and situationally dependent, the principle that prescribes it is only effective insofar as it defines a lower threshold below which control is so minimal to become irrelevant. Hence, the principle can be implemented minimally and maximally. Minimally, the implementation of this principle requires having a human *on* the loop able to understand the functioning of the system and its implications and with the ability to 'unplug' the system timely and effectively. Maximally, the principle requires individuals in charge of AI systems to combine technical, legal and ethical training to ensure that the decision *to let the system work* is informed by all relevant dimensions and not a mere vetting of the system.

Therefore, the principle does not admit *fire and forget* uses of AI, as it considers control as an element which can be modulated with respect to a rigorous risk assessment of unintended consequences, and related negative impact on national defence

and international stability. Where even lower levels of meaningful control cannot be complemented with these assessments, the use of AI systems is ethically unwarranted. It should be noted that the principle is best implemented when protocols for the attribution of responsibilities for united outcomes, misuses of AI and mistakes made by AI systems are in place alongside effective redressing and remedy processes. The attribution of responsibility hinges on the respect of transparency.

5.5 Reliable AI Systems

Defence organisation using AI systems must establish meaningful monitoring of the execution of the tasks delegated to AI. The monitoring should be adequate to the learning nature of the systems, and their lack of transparency, while remaining feasible in terms of resources, especially time, and hence computational feasibility.

AI has a poor shock response (robustness), and any slight alterations to inputs can degrade a model disproportionately (Rigaki & Elragal, 2017). Thus, deploying on AI for defence purposes could favour opponents (Brundage et al., 2018; Taddeo et al., 2019), if the system is not deployed according to procedure that envisage forms of monitoring and prompt intervention in case of mistakes or system degradation. This is why this principle prescribes monitoring of the systems throughout their deployment on top of having in place measures that verify and validate the systems and assess their robustness.

Monitoring may include new forms of procurement that envisage an active role of the defence institutions in the design and development process; in house design and development of AI models; use of data for system training and testing collected, curated and validated directly by the systems providers and maintained securely; mandatory forms of adversarial training with appropriate levels of refinement of AI models to test their robustness; sparring training of AI models; and monitoring the output of AI systems deployed in the wild with some form of in silico baseline model, as suggested by Taddeo et al. (2019).

As stressed in Sect. 2 of this article, AI systems are autonomous, self-learning agents interacting with the environment. Their behaviour depends as much on the inputs they are fed and interactions with other agents once deployed as it does on their design and training. Responsible uses of AI for defence purposes need to take into account the autonomous, dynamic and self-learning nature of AI systems, and start envisaging forms of monitoring that span from the design to the deployment stages.

6 Conclusion

As we mentioned at the beginning of this article, ethical principles for the use of AI in defence do not undermine international humanitarian laws. Rather, they offer guidance both with respect to what can be done post-compliance and with respect to those uses of AI in defence which international humanitarian laws do not address or do not address clearly. At the same time, the principles need to be logically

consistent with the broader ethical principles underpinning the wide set of uses of AI in our societies, like, for example, the OECD principles, and with the values shaping defence institutions and their role in democratic societies. For example, the US DIB states clearly in its documents that its principles rest on international humanitarian law, as well as on core values of the US Armed Forces. This consistency is important, for it ensures that despite the domain-dependent differences, ethical principles shaping the uses of AI in defence remain coherent with fundamental principles of our societies. This is crucial, for it will shape the trade-offs among the proposed principles that will have to be made from time to time and which will vary with the context of deployment.

Finally, we would like to conclude our analysis with a warning. These principles should not be followed as an algorithm, and they do not offer a set of instructions that, if followed slavishly, ensure ethically sound outcomes. They offer guidelines to spur and articulate ethical considerations with respect to the uses of AI in defence. To this end, it is a key that both humans making the decision to use AI and those executing these decisions are able to take into account the principles offered in this article, along with knowledge of legal and technical aspects of AI with the aim to reconcile different principles, interests and goals, without breaching fundamental values and rights of our societies.

Acknowledgements We are very grateful to Isaac Taylor for his work and comments on an early version of this article and to Rebecca Hogg and the participants of the 2020 Dstl AI Fest for their questions and comments, for they enabled us to improve several aspects of our analysis. We are responsible for any remaining mistakes.

Funding Mariarosaria Taddeo and Alexander Blanchard'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 contained in this paper cannot supersede any statutory or contractual requirements or liabilities and is offered without prejudice or commitment.

References

Acalvio Autonomous Deception. 2019. Acalvio. 2019. https://www.acalvio.com/.

Asaro, P. (2012). On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross, 94*(886), 687–709. https://doi.org/10.1017/S1816383112000768

BehavioSec: Continuous Authentication Through Behavioral Biometrics. 2019. BehavioSec. 2019. https://www.behaviosec.com/.

Boardman, Michael, and Fiona Butcher. 2019. An exploration of maintaining human control in AI enabled systems and the challenges of achieving it. STO-MP-IST-178.

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.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. ArXiv:1802.07228 *[Cs]*, February. http://arxiv.org/abs/1802.07228.

Brunstetter, D., & Braun, M. (2013). From Jus Ad Bellum to Jus Ad Vim: Recalibrating our understanding of the moral use of force. *Ethics & International Affairs, 27*(01), 87–106. https://doi.org/10.1017/S0892679412000792

DarkLight Offers First of Its Kind Artificial Intelligence to Enhance Cybersecurity Defenses. 2017. Business Wire. 26 July 2017. https://www.businesswire.com/news/home/20170726005117/en/DarkLight-Offers-Kind-Artificial-Intelligence-Enhance-Cybersecurity.

DeepLocker: How AI Can Power a Stealthy New Breed of Malware. 2018. *Security intelligence* (blog). 8 August 2018. https://securityintelligence.com/deeplocker-how-ai-can-power-a-stealthy-new-breed-of-malware/.

Department for Digital, Culture, Media & Sport. 2018. Data ethics framework. https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework.

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

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

Docherty, Bonnie. 2014. Shaking the foundations: The human rights implications of killer robots. Human Rights Watch. https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots.

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

Enemark, C. (2011). Drones over Pakistan: Secrecy, ethics, and counterinsurgency. *Asian Security, 7*(3), 218–237. https://doi.org/10.1080/14799855.2011.615082

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329. https://doi.org/10.1007/s11023-008-9113-7

Floridi, L. (2016a). Mature information societies—A matter of expectations. *Philosophy & Technology, 29*(1), 1–4. https://doi.org/10.1007/s13347-016-0214-6

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

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology, 31*(1), 1–8. https://doi.org/10.1007/s13347-018-0303-9

Floridi, Luciano, and Josh Cowls. 2019. A unified framework of five principles for AI in society. *Harvard Data Science Review*, June. https://doi.org/10.1162/99608f92.8cd550d1.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics, 26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fraga-Lamas, Paula, Tiago M. Fernández-Caramés, Manuel Suárez-Albela, Luis Castedo, and Miguel González-López. 2016. A review on Internet of things for defense and public safety. *Sensors (Basel, Switzerland)* 16 (10). https://doi.org/10.3390/s16101644.

Gavaghan, Colin, Alistair Knott, James Maclaurin, John Zerilli, and Joy Liddicoat. 2019. Government use of artificial intelligence in New Zealand, final report on phase 1 of the Law Foundation's Artificial Intelligence and Law in New Zealand Project. In. New Zealand Law Foundation: Wellington. https://www.cs.otago.ac.nz/research/ai/AI-Law/NZLF%20report.pdf.

International Telecommunications Union. 2017. Minimum requirements related to technical performance for IMT-2020 radio interface(s). 2017. https://www.itu.int/pub/R-REP-M.2410-2017.

Japanese Society for Artificial Intelligence [JSAI]. 2017. Ethical guidelines. http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics, 12*(2), 129–141. https://doi.org/10.1080/15027570.2013.818399

King, Tariq M., Jason Arbon, Dionny Santiago, David Adamo, Wendy Chin, and Ram Shanmugam. 2019. AI for testing today and tomorrow: Industry perspectives. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 81–88. Newark, CA, USA: IEEE. https://doi.org/10.1109/AITest.2019.000-3.

Kott, Alexander, Ananthram Swami, and Bruce J. West. 2017. The Internet of battle things. ArXiv:1712.08980 *[Cs]*, December. http://arxiv.org/abs/1712.08980.

Lysaght, Robert J., Regina Harris, and William Kelly. 1988. Artificial intelligence for command and control. ANALYTICS INC WILLOW GROVE PA. https://apps.dtic.mil/docs/citations/ADA229342.

McMahan, Jess. 2013. Forward. In *Who should die? The ethics of killing in war*, edited by Ryan Jenkins, Michael Robillard, and B J Strawser, ix–xiv. Oxford ; New York, NY: Oxford University Press.

Mirsky, Yisroel, Tom Mahler, Ilan Shelef, and Yuval Elovici. 2019. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. *ResearchGate*. https://www.researchgate.net/publication/330357848_CT-GAN_Malicious_Tampering_of_3D_Medical_Imagery_using_Deep_Learning/figures?lo=1.

Mökander, Jakob, and Luciano Floridi. 2021. Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, February. https://doi.org/10.1007/s11023-021-09557-8.

Nagel, T. (1972). "War and Massacre." Philosophy and Public Affairs, 1972, 1 (Winter): 123-144'. 1972. *American Behavioral Scientist, 15*(6), 951–951. https://doi.org/10.1177/000276427201500678.

NATO. 2020. NATO 2030: United for a new era. Brussels. https://www.nato.int/nato_static_fl2014/assets/pdf/2020/12/pdf/201201-Reflection-Group-Final-Report-Uni.pdf.

O'Connell, Mary Ellen. 2014. The American way of bombing: How legal and ethical norms change. In, edited by Matthew Evangelista and Henry Shue. Ithaca: Cornel University Press.

Rigaki, Maria, and Ahmed Elragal. 2017. Adversarial deep learning against intrusion detection classifiers. In, 14.

Roberts, Huw, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2020. The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & SOCIETY*, June. https://doi.org/10.1007/s00146-020-00992-2.

Schubert, Johan, Joel Brynielsson, Mattias Nilsson, and Peter Svenmarck. 2018. Artificial intelligence for decision support in command and control systems, 15.

Sharkey, A. (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, N. (2010). Saying "No!" to lethal autonomous targeting. *Journal of Military Ethics, 9*(4), 369–383. https://doi.org/10.1080/15027570.2010.537903

Sharkey, N. (2012a). Killing made easy: From joysticks to politics. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 111–128). MIT Press.

Sharkey, N. E. (2012b). The evitability of autonomous robot warfare. *International Review of the Red Cross, 94*(886), 787–799. https://doi.org/10.1017/S1816383112000732

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs, 30*(1), 93–116. https://doi.org/10.1017/S0892679415000647

Taddeo, M. (2012a). Information warfare: A philosophical perspective. *Philosophy and Technology, 25*(1), 105–120.

Taddeo, Mariarosaria. 2012. An analysis for a just cyber warfare. In *Fourth International Conference of Cyber Conflict*. NATO CCD COE and IEEE Publication.

Taddeo, M. (2013). Cyber security and individual rights, striking the right balance. *Philosophy & Technology, 26*(4), 353–356. https://doi.org/10.1007/s13347-013-0140-9

Taddeo, Mariarosaria. 2014a. Just information warfare. *Topoi*, April, 1–12. https://doi.org/10.1007/s11245-014-9245-8.

Taddeo, Mariarosaria. 2014b. The struggle between liberties and authorities in the information age. *Science and Engineering Ethics*, September, 1–14. https://doi.org/10.1007/s11948-014-9586-0.

Taddeo, M. (2017). The limits of deterrence theory in cyberspace. *Philosophy & Technology*. https://doi.org/10.1007/s13347-017-0290-2

Taddeo, Mariarosaria. 2019a. The challenges of cyber deterrence. In *The 2018 Yearbook of the Digital Ethics Lab*, edited by Carl Öhman and David Watson, 85–103. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-17152-0_7.

Taddeo, M. (2019b). Three ethical challenges of applications of artificial intelligence in cybersecurity. *Minds and Machines, 29*(2), 187–191. https://doi.org/10.1007/s11023-019-09504-8

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature, 556*(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6

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

Tamburrini, G. (2016). On banning autonomous weapons systems: From deontological to wide consequentialist reasons. In B. Nehal, S. Beck, R. Geiβ, H.-Y. Liu, & C. Kreβ (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 122–142). Cambridge University Press.

The UK and International Humanitarian Law 2018. n.d. Accessed 1 November 2020. https://www.gov.uk/government/publications/international-humanitarian-law-and-the-uk-government/uk-and-international-humanitarian-law-2018.

US Army. 2017. Robotic and autonomous systems strategy. https://www.tradoc.army.mil/Portals/14/Documents/RAS_Strategy.pdf.

Yang, Guang-Zhong, Jim Bellingham, Pierre E. Dupont, Peer Fischer, Luciano Floridi, Robert Full, Neil Jacobstein, et al. 2018. The grand challenges of *Science Robotics*. *Science Robotics* 3 (14): eaar7650. https://doi.org/10.1126/scirobotics.aar7650.

Zhuge, Jianwei, Thorsten Holz, Xinhui Han, Chengyu Song, and Wei Zou. 2007. Collecting autonomous spreading malware using high-interaction honeypots. In *Information and communications security*, edited by Sihan Qing, Hideki Imai, and Guilin Wang, 438–51. Lecture Notes in Computer Science. Springer Berlin Heidelberg.