FEATURE

Taming the Killer Robot

Toward a Set of Ethical Principles for Military Artificial Intelligence

Dr. Jai C. Galliott

Dr. Massimiliano L. Cappuccio

Dr. Austin Wyatt

Abstract

Both corporate leaders and military commanders turn to ethical principle sets when they search for guidance concerning moral decision making and best practice. In this article, after reviewing several such sets intended to guide the responsible development and deployment of artificial intelligence and autonomous systems in the civilian domain, we propose a series of 11 positive ethical principles to be embedded in the design of autonomous and intelligent technologies used by armed forces. In addition to guiding their research and development, these principles can enhance the capability of the armed forces to make ethical decisions in conflict and operations. We examine the general limitations of principle sets, refuting the charge that such ethical theorizing is a misguided enterprise and critically addressing the proposed ban on military applications of artificial intelligence and autonomous weapon systems.

Regulating the use of artificial intelligence (AI) to make it safe and compliant with ethical standards recently became a public concern and a global priority. One of the controversies most intensely debated by technology ethicists remains the military application of AI, which includes—but is not limited to—autonomous weapon systems (AWS), i.e., machines designed to independently search for, select, and engage targets.[1] A universal ban on these machines has been advocated by those who believe that the conditions necessary to use AWS ethically either are impossible to devise or cannot realistically be met in practice.[2] Articulating an alternative proposal, we argue that the conditions for ethically using military applications of AI can be conceptually specified as clearly as those relevant to similar nonmilitary technologies; that the decisional processes (including the public discussion) and the research efforts (including the transfer from civilian to military industry) necessary to practically meet such conditions would be hindered by a pre-emptive ban on AWS; and that any such unconditional prohibition would solicit the very same deregulation and uncontrolled proliferation that it was supposed to prevent. To avoid that the prophecy fulfills itself, we recommend that each instance of design, development, and deployment of AWS should be internationally regulated by legal and ethical standards. Compared to an indiscriminate ban, this approach would be more politically efficacious and authentically moral.[3]

What distinctively characterizes our proposal is the suggestion that any normative framework for AWS should make parallel with the codes of practice already established to regulate civilian technologies, such as commercial drones and autonomous cars. That is because, aside from obvious specificities, military and civilian applications of AI face comparable ethical challenges and must align with fundamentally analogous shared values and societal expectations.

While several ethical standards have already been implemented in the domain of civilian uses of AI, governments have yet to agree a shared ethical framework to regulate military AI. The first step in this direction is represented by the five self-regulation principles with which the US Department of Defense (DOD) commits to ensuring that military AIs incorporate ethical characteristics:

1. *Responsible* (informed by "appropriate levels of judgment and care");
2. *Equitable* (minimizing "unintended bias in AI capabilities");
3. *Traceable* (using "transparent and auditable methodologies, data sources, and design procedure and documentation");
4. *Reliable* (fulfilling rigorous "safety, security, and effectiveness" standards); and
5. *Governable* (neutralizable when systems "demonstrate unintended behavior").[4]

The commitment of the DOD, which is expected to inform all the relevant policies and practices by the Pentagon and its affiliated entities, has not corresponded to equivalent efforts by private companies that work as contractors for the military industry.[5] This hesitation is well exemplified by Project Maven, which was aborted after an initial collaboration agreement between Google and the DOD.[6] The project aimed at using Google's voluminous video dataset to train an automated visual recognition system to be applied by military drones for purposes of reconnaissance and targeting.

Controversies arising among Google's staff, and subsequent public pressure, induced Google to divest itself of its contract with the DOD. Internal discontent was not quelled even when Google chief executive released a public set of "guiding principles" that prevent Google from researching "offensive" applications of AI without restricting the company's intention to develop military technologies with the Pentagon.[7]

Due to the virtual impossibility of establishing in advance which applications of AI can or cannot be weaponized (Project Maven itself was not expected to serve only offensive purposes), this code of practice appeared remarkably ambiguous: on the one hand, it does not say or imply that military research is inherently immoral (in fact, Google publicly admitted its eagerness to sign new contracts with the Pentagon);[8] on the other hand, it never commits to a rigid series of requirements that would, by their fulfillment, imply that the firm considered participation in the military applications of AI to be appropriate and ethically just.

The series of corporate ethics frameworks that sprung up in the wake of Google's decision indicate a similar reluctance to specify stringent standards under which private firms would consider collaboration on military AI research to be ethically permissible. This hesitancy does not stem from an aversion to ethical standards upon research and development (R&D), as both civilian firms and military institutions acknowledge the need for codes of ethics, but from obscurity regarding whether the same ethical and legal norms should govern civilian and military research into AI.

If AI technologies can have dual uses (civilian and military), why then cannot the related AI ethical principles have a dual implementation? Certainly, civilian and military research are governed by different laws and ethical protocols, but their efforts toward the development of trusted autonomous technologies involve analogous challenges (e.g., transparency, predictability), methods (e.g., data acquisition, machine learning), and desired functionalities (e.g., computer vision) to complete very similar tasks (e.g., autonomous navigation and context-adaptive action selection) while managing the same kinds of risks (e.g., fatal misrecognitions, biased datasets, unfair decisional criteria, technological dependency, etc.).

Thus, far from being exceptional, the ethical concerns arising from the military applications of AI should be proportioned to those associated with mass-scale civilian applications of AI (e.g., autonomous driving, predictive policing).[9]

However, the debate on military AI tends to be more divisive than the equivalent debate on the civilian applications of AI. Unfortunately, this could hinder the effective transfer of ethical guidelines from the civilian industry to the military. The divisiveness depends much more on the *public perception of* the differences between civilian and military ethical standards than their objective magnitude. Three contextual elements contribute to amplify the perception of these differences: (1) the inhomogeneity of the criteria adopted by different nations to define autonomy and its related responsibilities, which reflect a diversity in political visions and strategic priorities;[10] (2) the tendency of the general public and media to represent military AI technologies as more unreliable and unsafe than the homologous civilian technologies;[11] and (3) the movements of public opinion that advocate a pre-emptive ban on AWS rather than a discussion on the possibility of ethically regulating their use.[12]

We shall focus on the third element because it has the power to exacerbate the other two: paradoxically, the proposal for a pre-emptive ban has greatly contributed to delaying constructive discussions into military AI ethics,[13] discouraging any effort to establish appropriate guidelines for developing and deploying ethically compliant AWS.

The Ban on Killer Robots and Its Paradoxes

Let us consider, for example, the proposal to create a new international treaty under Additional Protocol 1 of the Geneva Conventions, a position illustrated by the *Open Letter to the United Nations Convention on Certain Conventional Weapons*, which raised the concern that "killer robots" may pervade warfare to the point of making armed conflict more frequent, inhumane, and uncontrollable.

This scenario is not corroborated by previous cases, but it is motivated by two worries. The first is that deferring tactical decisions to machines may favor the moral deskilling of military personnel[14] and desensitize military command to the human costs of war,[15] hence partly exculpating them and transforming AWS into scapegoats for human errors and advancing undue overconfidence, with potential to yield undesirable consequences such as bringing conflicts into urban areas. The second being that AI will never be accurate or reliable enough to apply lethal force *proportionately* (being sufficiently selective in acquiring the relevant targets) and *discriminately* (distinguishing enemy combatants from civilians), which would prevent the deploying force from respecting international humanitarian law and satisfying the basic requirements of classical just war theory (*in bello*).[16]

Both worries are reasonable in themselves and need to be addressed.[17] However, such concerns could logically motivate a complete ban on AWS only if the following implicit assumption holds: that military AI will never be able to incorporate the required ethical principles and guidelines because the military developers cannot anticipate the errors that the machines may make. Thus, what truly concerns the letter's signatories is *neither the practical limitations of technology per se* (such limitations are never leveraged to motivate, for example, an *unconditional ban* of civilian autonomous vehicles) nor the *theoretical impossibility of producing appropriate regulations* (as the letter never denies that appropriate norms and protocols could be established in principle).

Rather, the real concern seems to be the *alleged structural unreadiness or unwillingness of nations, and specifically their armed forces, to effectively establish and enforce such regulations*: the worry here is that the haste to exploit the formidable advantages offered by AWS would obfuscate command's ability to recognize and counter AWS's inherent risks.[18] Among the petitioners, some AI experts fear that military decision makers are too optimistic, naïve, or distracted to recognize that machine-learning techniques are opaque and tend to produce unpredictable and unexplainable patterns of behavior.[19]

This fear hardly reflects how innovative technological capabilities are thoroughly tested before being adopted by the armed forces of democratic countries, following compliance protocols and safety standards that are dictated by the very same tactical needs for which these capabilities are developed in the first place, therefore, are not less stringent, demanding, and transparent than their civilian homologues.[20] The sustained efforts made by researchers to minimize unpredictable and unexplainable behaviors are too often overshadowed by the tendency of media (e.g., Samuel Gibbs)[21] to depict AWS with science-fiction clichés and portray the armed forces as mesmerized by the inevitable ascent of "Terminator"-like robots. In this hypothetical "sorcerer's apprentice" scenario, military decision makers are both tantalized and overwhelmed by the uncontrollable complexity and intrinsic subversiveness of their technological creations, consumed by an implausible combination of hubristic ambitions and hypnotic passiveness.

This scenario would be less implausible if all the tactically advantageous applications of military AI happened to be also ethically undesirable, as this tension could compel the armed forces to prioritize military over moral values.[22] The truth is that military applications of AI can improve selectivity, accuracy, accountability, and traceability; hence, helping to minimize collateral damages and reducing casualties—offering an important opportunity to make armed conflicts less unethical and inhumane.[23] Some applications of autonomy are already desirable insofar as they offer acceptable trade-offs between risks and benefits in the tactical and

operational dimensions, where benefits must be evaluated against critical stakes and risks can be mitigated but never entirely eliminated.[24]

The request to completely ban AWS shoehorns the debate on military AI into a false dichotomy: either the development of all autonomous weapons is put on hold due to its unacceptable risks, or it will be continued ignoring such risks. This false dichotomy, instead of offering a *solution* to said risks, becomes part of the ideological *problem* that deters military contractors from developing ethical principles for AI. The problem, we believe, consists in aprioristically rejecting the possibility that armed forces could ever envision and enforce appropriate AI regulations. Such a position suffers from at least four flaws:

> *a) An ethical flaw*: dismissing the very possibility that military AIs could be designed ethically delays the development of ethical military applications, making warfare more cruel, unjust, and exposed to greater threats for human dignity. AWS capabilities are not only morally desirable for utilitarianist reasons but also deontologically obligatory in any scenarios in which the deterrence power of an assured military response, the armed forces' capability to deliver such a response with sufficient "speed, precision, efficiency," the "distribution of responsibility and guilt" guaranteed by an impersonal lethal system, the upholding of human dignity in death, and the sustainable management of resources represent non-negotiable ethical imperatives.*, [25] Also, because some military applications of AI can, in principle, and do, in practice, "satisfy the requirements of fairness in the re-distribution of risk," "there is a public responsibility to regulate [not ban] killer robots' design and manufacture."[26]

* Such scenarios include "saving your village by activating a robot to kill invading enemies who would inflict great indignity on your village, using a suicide robot to save yourself from a less dignified death at enemy hands, using a robotic drone to kill someone otherwise not accessible in order to restore dignity to someone this person killed and to his family, and using a robot to kill someone who needs killing, but the killing of whom by a human executioner would soil the executioner's dignity. I conclude that what matters in rightful killing isn't necessarily that it be under the direct control of a human, but that it be under the control of morality." Duncan MacIntosh, "Fire and Forget: A Moral Defense of the Use of Autonomous Weapons Systems in War and Peace," in *Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare*, ed. Jai Galliott, Duncan MacIntosh, and Jens David Ohlin (New York: Oxford University Press, 2021).

b) A methodological (scientific) flaw: presuming that military AI are irremediably untrustworthy prevents researchers from collecting the data and developing the scientific models necessary to assess and increase the trustworthiness of all autonomous systems, including both the military and the civilian ones. Prohibiting the research on AWS not only deprives states of the ethical advantages of military AI but also confounds the ethical risks of the corresponding civilian applications, overshadowing the extent and utility of dual-use technologies.

Addressing this conundrum, a recent study indicates that the roboticists supporting the ban often lack proper awareness of "the real-world dual-use potential of their creations," which is why many civilian developers are "not well suited for analyzing how an autonomous system designed for a humanitarian emergency may have capabilities that are directly transferrable to a conflict zone."[27] The disconnection of civilian and military research is not just "a matter of willful ignorance of technology developers but rather a void created, intentionally or unintentionally, by policymakers." This void, which arises from "the lack of agreement on norms and regulations at the international level about autonomous systems," inevitably discourages any effort to design and develop more ethical military AI.

c) A communicative (political) flaw: obstructing constructive discussions about the ethical applications of military AI delays the diplomatic and legislative processes that may lead to establishing the relevant international treaty. Part of the problem is that the pro-ban campaign "is modelled after previous humanitarian disarmament successes and not tailored to the specifics [of AWS]", whose irreducible novelty must be researched and acknowledged without prejudices.[28] Unlike other armament systems that have been successfully banned (e.g., blinding laser weapons and antipersonnel landmines), AWS are mostly unmanned, lethal, military variations of manned, nonlethal, civilian or dual-use technologies.[29] The distinction between one of these technologies and its lethal counterpart, from an AI

perspective, is extremely thin or imperceptible, which is why both state and nonstate actors could easily find ways to work around a general ban, cheating "the treaty regime by developing autonomous software for weapon systems that are normally manned"[30] as well as weaponizing algorithms developed for civilian applications.

The very definition of the autonomous technology to be banned is disappointingly vague and has led some to conflate AWS with military applications of AI tout court. The most cogent definitions of the term *autonomous weapons system* base their classification on the extent to which the system has control over its critical functions independent of a human, for which AI is a vital enabler. Despite seeming agreement on its broad strokes, definitions of AWS begin to differ when it comes to the crucial technical details and standards upon which any effective ban would have to rely,[31] so that a ban on AWS would predictably end up forbidding all autonomous technologies or with no technology at all.

Not only does this lack of agreement make the ban *formally empty* (its scope is either over- or underspecified, making the prohibition inapplicable) and *hardly enforceable* (due to the difficulties arbiters would face in verifying whether a weapon system is "autonomous" without being granted high levels of technical access by the deploying military),[32] but it also makes the ban *politically counterproductive*: instead of preventing proliferation, this vagueness might encourage national armed forces and private militias to give rise to "highly secretive programmes" to develop AWS disguised as civilian or nonlethal technologies,[33] which would be both procedurally and morally unacceptable.

A pre-emptive ban would limit the international community's propensity and ability to effectively develop and systematically apply definitions, standards, norms, and practices, at a time when states urgently need to agree and enforce official guidelines to rigorously review and test the

design concept, general functionality, and tactical viability of any AWS before they are deployed into actual combat. For example, only guidelines of this kind could help states agree how AWS should interact with each other, ensuring that no unintended escalation results from the random interaction of two states' autonomous weapons.[34]

d) A governance (motivational) flaw: the ban incentivizes exactly the irresponsible and risky decisions that it is meant to prevent. Military contractors are deontologically committed and legally obligated to create trusted autonomous systems, as testified by their efforts to develop trust marks for AWS and protocols to objectively assess their autonomy levels.[35] However, while the developers and contractors officially working for the armed forces of democratic states have an obvious interest and the strongest motivations to identify and discontinue the untrustworthy applications of AWS, those hired by rogue militias or non-state paramilitary groups secretly operating around an international ban might not feel the same moral urgency.[36]

This point seems largely ignored by the pro-ban campaigners, who reductively conflate a practical, concrete problem (the immaturity of current technology) with an intangible fear (the alleged unreadiness or unwillingness of armed forces to comply with ethical standards). Not only is this negative assumption about the armed forces' motivations unjustified, but also the conflation itself is dangerous.

To be ethically justified, the use of autonomy in warfare must fulfill three necessary conditions: (a) AWS's conceptualization and design must be informed by sound moral goals and values (e.g., collateral damage minimization); (b) AWS's intended functioning must be supported by sufficiently sophisticated technological solutions to keep up with the moral expectations of the designers (e.g., advanced visual recognition systems); and (c) decisions to deploy AWS must comply with the relevant deontic norms that govern legitimate just war operations. These norms are informed not only by the relevant military doctrines (e.g.,

rules of engagement, just war theory), but also by ethical considerations that specifically apply to autonomy, reflecting the values that inform their design. Any ethical use of AWS must fulfill these three conditions, of which technological maturity is but one. A ban cannot differentiate between these three requirements because it operates under the assumption that, when they are pressed to meet the technological requirement, military decision makers inevitably discount the other two requirements. Such a conflation is simplistic because, within an appropriate regulative framework, these three requirements are equally important and can be addressed separately.

Even if specific concerns about the maturity of this or that technology may be justified, the range of military applications of AI is so broad and innovative at so many different levels, and including so many degrees of potential integration with human supervision, that giving the same one-size-fits-all moral judgment to all the possible uses of AI in warfare risks being not only simplistic but also demotivating for those who are making genuine efforts to make warfare less inhumane.[37] It is reasonable to worry that technological innovation may at times outpace the efforts to govern it, but, instead of a withdrawal, this problem should motivate even more systematic efforts to clarify the normative principles under which AWS should operate and ensure that the relevant governance frameworks are constantly updated based on empirical evidence and ethical awareness.[38]

To avoid these four flaws, we propose to replace the pre-emptive ban with a set of ethical principles. While specifically designed for the military, this set is importantly informed by how similar principles are implemented by the civilian AI industry. Without overlooking the theoretical limitations of the principlist approach to applied ethics,[39] we will argue that an appropriately designed set of principles may, if nothing else, avoid the most obvious ethical violations in wartime and guide the development of a "minimally just AI" (MinAI) framework[40] to inform the design of any AWS.

Technological Society's Love Affair with Ethical Principles

Once it was publicly revealed that Google was collaborating with the Pentagon on Project Maven, Google's work on AI was to be urgently made socially responsible by adhering to ethical principles that included a commitment to "be socially beneficial" and to "avoid, creating or reinforcing unfair bias." Others dictate that their wares "be built and tested for safety," "incorporate privacy design principles," "uphold high standards of scientific excellence," and "be accountable to people."[41] While signaling that the company's intention was not to divest itself of military contracts, Google's principles set also includes a section titled, "AI applications we will not pursue'," which directly mentions "weapons and other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people."[42] It is far from clear who would ultimately maintain responsibility for the implementation of the principles, as the Google ethics board, which existed barely more than one week, has been disbanded owing to public discord over its members. This controversy reopened old divisions within the company, with no replacement board or mechanism having since been named.

Table 1. "Popular Tech-Ethics Principles by Comparison" or similar

US DOD	Microsoft	Future of Life Institute		European Commission	Salesforce
Equitable	Fairness	Safety	Liberty & Privacy	Human agency & oversight	Beneficence
	Inclusiveness	Failure Transparency	Shared Benefits	Robustness & safety	Human value alignment
Reliable	Reliability & Safety	Judicial Transparency	Shared Prosperity	Privacy & data governance	Open debate between science & policy
Traceable	Transparency	Responsibility	Human Control	Transparency	Cooperation, trust, & transparency in systems and among the AI community
Governable	Privacy & security	Value alignment	Non-subversion	Diversity, non-discrimination, & fairness	Safety & responsibility
Responsible	Accountability	Human values	AI arms race	Social & environmental wellbeing	-
	Personal privacy			Accountability	

Google is just one example of a company resorting to ethics principles in the face of technological challenges. Despite the dissolution of its AI ethics board, a wave of ethics principles has swept Silicon Valley since, as those holding interests

in AI come to realize the potentially controversial nature of its applications and the need to curb unintended dual- or suspicious uses that may impact their reputation. AI ethics has therefore come to be of interest across a number of civil sectors and types of institutions, ranging from small- to large-scale developers of technology eager to generate their own ethical principles, professional bodies whose codes of ethics are aimed at influencing technical practitioners through standards setting, and monitoring bodies such as research institutes, government agencies, and leading individual researchers across disciplines whose work aims to add technical or conceptual depth to AI.

This wave, preceding both Google and the DOD's efforts, was not limited solely to technology corporations.[43] At around the same time as the Future of Life Institute (FLI) released its "Asilomar AI" principles,[44] the US Association for Computing Machinery (ACM) communicated a set of seven principles focused on algorithmic opacity and its connection to responsibility attribution.[45] In 2017 alone, several other stakeholder groups and organizations published additional principles sets, including the Japanese Society for AI's Ethical Guidelines;[46] a set of recommendations from the Université de Montréal, entitled the Montréal Declaration on Responsible AI;[47] and the IEEE General Principles of Ethical Autonomous and Intelligent Systems.[48] This proliferation of principles has continued into the following years with the Partnership on AI producing a set of "tenets" that its members agree to uphold,[49] and the UK House of Lords suggesting five principles for a cross-sector but nonmilitary AI code, which could be adopted internationally and was based on the evidence of some 200 experts.[50]

Google's interest in presenting a forward-leaning perspective on ethical AI use influenced this wave, inspiring the Microsoft AI principles,[51] which revolve around designing AI to be "trustworthy" and emphasize values like accountability, reliability, and safety. Trustworthiness and accountability were also central to the views of the European Commission,[52] which announced a list of "seven essentials for achieving trustworthy AI" and highlighted the importance of robustness and safety, transparency, and accountability for the development of ethical AI.

Notably, value alignment and respect of the democratic processes of liberal society are common across each of these frameworks, which emphasize that AI research must further the empowerment of people through technological tools used globally on a daily basis: e.g., 10 of the 13 principles developed by FLI[53] were directly related to the preservation of human rights, privacy and dignity; similarly, four of the five principles endorsed by Technology & Products of Salesforce centered on this theme.[54] Table 1 contains a list of ethical AI principles from a selection of corporate and civil society organizations.

These principles are united by having been formulated by authoritative experts and leaders and endorsed by their stakeholder communities. Our normative framework for military AI is inspired by the ethical principles developed in the civilian sphere, from which we draw key themes (acceptance, control, transparency, fairness, and safety) that we seek to operationalize through a consistent set of principles.

Ethical Principles for Military AI

Amid growing fears of biased and weaponized AI,[55] and the faulty logic of the reductionist pre-emptive ban position earlier problematized, armed forces need to systematically adopt ethical AI principles to guide their responsible approach to the development and acquisition of AWS. It is plausible that such principles will be initially adopted by a closely aligned block of militaries that already participate in joint programs and share comparable military cultures (e.g., the Five Eyes intelligence alliance or an EU-led military alliance). Our set of 11 principles is inspired by the above-detailed civilian principles while aligning with and further developing the five principles recommended by the DOD, emphasizing value-related questions in the design and development process.

(I) **Be socially accepted and politically legitimized.**

Unlike any other armament system, the autonomous behavior and selective processes of AI-based weapons can be infused with very specific rules and criteria, which, in turn, can and must be carefully customized to reflect the civil society's values, principles, and norms. Moreover, an AI-based weapon can directly respond to the needs and interests of democratic nations because its specific behaviors and functioning settings can be modified over time (during a mission and even seconds before impacting the target) to adjust to the decisions made by legitimate leaders and institutions.

Military ethical AI principles must be developed and regularly reviewed for impact by a panel of independent experts, respecting the rules-based system of international order. This is important considering that the barriers for acquiring some AWS are relative-

ly low: nowadays, the cheapest—thus more primitive and unsafe—models can be afforded even by terrorist groups without techno-scientific edge.

Social acceptance, generally recognized as one of the core principles for responsibly conducting scientific research, is a key factor in determining the legitimacy of military technology. For a military system to achieve social acceptance, its design, development, and deployment must reflect the ethical standards that frame a military's role in their society.[56] Unlike unrecognized paramilitary groups and rogue private militias, the status of national armed forces operating under rule of law legitimizes their conduct through adherence to international law and institutional order and remains beholden to states via a public mandate that must be maintained.

As the institutions of the state are an expression of a social contract,[57] the very existence of military institutions and the justification of their authority—including the power to develop and deploy AWS—must derive from such a contract. This implies that military AI should not be used to undermine the culture, the form of political organization, and the values (e.g., civil rights) that distinctively characterize the community. Take, for example, the reluctance of the Singapore Armed Forces to pursue weapon systems that would be perceived as overly aggressive by their neighbors. This stems primarily from the civilian government's defensive geopolitical stance, rather than military purpose.[58]

(II) **Be nationally compliant.**

Military autonomous technologies must comply with relevant national standards, laws, and policies. This is similar to how—with specific regard to safety regulations for small- to mid-sized remote piloted aircraft—the Australian Defence Force defers to the

authority of the Civil Aviation Safety Agency when operating remote piloted aircraft within Australian airspace to prevent serious disconnect between the way the military and the civilian authorities use remote piloted aircraft. Also, in a liberal society with a strong civil-military relationship, the internal use of military AI must follow the same national privacy and data-sharing regulations as law enforcement agencies, with a strong emphasis against the unjust gathering or storing of civilian information. While agreeing to a common set of rules and understanding, like other elements of the public and private sectors, the military must, at times, be authorized to override the limitations that apply to business and politics, e.g., during international operations that involve a distinct regulatory regime (such as international humanitarian law).

(III) **Be interoperable and mutually recognizable.**

Interoperability is vital because it grants allied armed forces superior coordination and greater cohesion. Military AI systems must be mutually recognizable among allied forces, be able to communicate with one-another when required, and allow for cross-training of allied officers or specialists. The principle of mutual recognition allows them to develop shared strategic visions, operational protocols, and tactics involving AI-based systems and practice them during joint operations. For example, the armed forces of the Five Eyes community are already conducting joint exercises (e.g., Autonomous Warrior 2019 and 2020)[59] to better understand how to incorporate autonomous agents into future operations. The shared commitment to the Law of Armed Conflict[60] is one of the key nontechnological preconditions to establish interoperability among allied armed forces.

(IV) **Be justified and explainable.**

As with any political decision to use force, the deployment of military AI must only occur in a just conflict for clearly articulable reasons. The responsibility of armed forces to adhere to these requirements is an important part of their social legitimacy and is often tied closely together with calls for greater transparency in strategic decision making. According to traditional doctrines of just war (*jus ad bellum*), "just conflicts" are those that meet six necessary conditions: right authority, right intention, reasonable hope, proportionality, and last resort.[61] Further, the use of weapons during just conflicts must be motivated by comprehensible good reasons.[62] Consequently, all decisions involving the use of weapon systems during a conflict, including the decisions autonomously made by AIs during operations, must be justified and communicable to other rational agents in an intelligible and transparent manner. Decisions made by AIs must result from algorithmic processes that can be analytically scrutinized and understood by those providing technical, operational, or political oversight. Armed forces must maintain a system of control that allows humans to exert at least an indirect, asynchronous influence over the decisions made by AWS, which, in turn, even when operating independently of human oversight, need to implement at least three criteria in maintaining: (1) a record of reliability; (2) a system to enable the comprehensibility of previous machine behavior and the predictability of its future behavior; and (3) a means to record human input to code and other system operations, such as the input of training data, to ensure data provenance and the accountability of particular AI applications.

These requirements are directly comparable to the role of transparency, accountability, and human responsibility in civilian versions of ethical AI: for example, technology firms are accountable to their shareholders and the customers for ensuring that

their product operates in an explainable manner.

(V) **Operate within a system of human control and accountability.**

Military AI systems must be designed and operated in such a way that all persons involved in their deployment and functioning are accountable for the work they conduct in this capacity. It must always be possible to evaluate the morality and legality of the decision to use AWS, reviewing the chain of responsibilities and assessing the risks stemming from deployment within the overarching system of human control in which autonomous systems operate (e.g., the Law of Armed Conflict, targeting doctrine, weapons review procedures, etc.). This requires building a positive feedback loop between users and developers to facilitate the required technological enhancements and improve ethical/operational outcomes,[63] without necessarily calling for a human to be in or on the loop at the execution of lethal action.[64]

Accountability and reliability represent crucial concerns for civilian developers, especially corporations that are held accountable for the functioning of their products. For example, consider the legal ramifications of a self-driving vehicle hitting a pedestrian: AI systems can malfunction, but they cannot be held responsible for the decisions that determined their immoral use, due to the impossibility of meaningfully apportioning blame to nonhuman entities.[65] The well-known "black box" and opacity problems of AI, which describe the possibility of unpredictable outcomes, should be treated as calculated risks deliberately taken on by both human users and makers.[66] Similarly, the opacities of AWS must ultimately be considered responsibilities for which military commanders and developers are to some extent accountable when they address uncertain and risky scenarios.

(VI) **Be built and certified safe and secure.**

As with their civilian counterparts, certification for the safe, secure, and reliable build and operation of military AI systems is needed for the prevention of harm and mitigation of risk to civilians and military personnel. In the civilian sphere, sophisticated electronics are inspected, certified and given manufacturer's warranties to assure that the product is of an expected standard. This requirement is backed by national and international legislation and is often subject to criteria imposed by multilateral trade deals.

The safety and certification regime for military autonomous systems should preferably be based upon international standards imposed by a supranational entity (e.g., the United Nations) or under international law. An alternative would be for allied militaries to jointly develop criteria for safety and security to evaluate the newly developed systems.

(VII) **Be ethical by design.**

Moral values and norms are to be carefully considered during the design phase of new military technologies.[67] An early review of potential ethical dilemmas should be undertaken from a design and engineering perspective to ensure that sufficient conceptual and empirical study is undertaken before the system (due to unnoticed imperfections or other causes) has any chance to engender strategic, economical, or humanitarian drawbacks. This principle corresponds to the corporate attention to embedding inclusiveness, diversity, and other customer values in product design. Similar analyses should be conducted by military developers relying on a value-sensitive design framework or similar methodology.[68] This would encourage the makers to emphasize the ethical significance of design values like adaptiv-

ity (the capability to adjust to context) and selectivity (the capability to discriminate between targets), which allow AWS to deliver better ethical outcomes in comparison with traditional weapon systems.

(VIII) **Avoid unjust bias.**

Military applications of AI are, just like their civilian counterparts, only as accurate as the dataset upon which they are trained. A core nontechnological barrier to the ethical use of AI, therefore, is the risk of unfair biases being injected into a military AI due to the nature of the dataset it is drawing upon. Such biases could deliberately or inadvertently reflect prejudices and discriminatory attitudes embedded in the maker's decisional patterns or the user's preferences, but they can also accidentally result from improperly selected and calibrated training datasets. A growing body of literature examines the impact of data-based biases in civilian AI systems,[69] focusing on biases toward individuals with darker skin,[70] gender-based discriminations induced by historical hiring patterns,[71] and discriminations in criminal risk assessments.[72] Recognizing unfair biases is not always simple or value-neutral as they are often implicit, socio-culturally connotated, and open to interpretation.

Armed forces must recognize unjust data biases that can impact on targeting decisions and identification of combatants, as highlighted by the unreliability of facial recognition software trained on databases of light-skinned individuals; similar training flaws could perpetuate harm to civilians preventing the accurate analysis of information about race, gender, sexual orientation, disability, ethnicity, or age. Dataset's verification may require targeting and surveillance protocols that focus exclusively on traits that are pertinent to the mission objectives. Ensuring that unjust data biases do not affect military AI sys-

tems would clearly benefit from further research on culture-specific human-machine interaction practices and algorithmic testing.

(IX) **Allow military personnel to flourish**

Despite the centrality of technology, soldiers' wellbeing and dignity must remain key principles in any development decision. As with private firms, training and retaining skilled and experienced people remains an ongoing priority. The development of AI must accommodate the continued capacity for military personnel to flourish mentally, emotionally, and economically. This includes preventing the deskilling of humans[73] and the improper transferring of authority and leadership roles to machines, which may prevent humans from expressing their freedom and potential.[74] This calls for further investigations on how humans tolerate AI, particularly where AI priorities compete with human values.

As the deployment of military AI can deeply affect the experiences of soldiers and veterans, its design should operate to preserve the dignity and the integrity of all the humans involved in all relevant circumstances (including those faced with the prospect of death); minimize any risk of psychological harm to soldiers, including post-traumatic stress disorder, burn-out, and moral injury (which often arises when soldiers have to transgress their moral beliefs); and prevent the risk of technological alienation (a neurotic condition resulting from the frustrating experience of serving a machine).[75]

(X) **Be sensitive to the environment**

Military applications of AI must minimize their impact on the environment. Protecting nature from unnecessary harm is enshrined in the laws of war and deserves equal importance in our conceptions

of ethically just deployments of AI. Preserving the conditions for human life and protecting important cultural objects through the appropriate use of AI are ethically desirable: the availability of natural resources is important to later-arriving personnel, and the support of local populaces is critical to lasting peace.[76] Sensitivity toward the environment can be found among the proclaimed values of major technology companies (including Google, IBM, and Samsung), and both the Australian and the US Departments of Defence maintain policies that identify environmental sustainability as a strategic interest. Cultural relics and sites of historical importance should further be protected from any engagement by AWS, by training the AI to recognize specific international protective symbols (like the UN Blue Shield for cultural assets worthy of protection).[77]

(XI) **Be malfunction-ready.**

Military AI and AWS must remain safe and controllable even when they behave in unexpected ways. Similar concerns encourage civilian roboticists to design systems that can "fail gracefully" in the event of malfunction. This capability is essential in the case of military systems, as designing them to malfunction safely in the field is arguably more important than attempting to make them failure-proof. The Oerlikon friendly fire incident in 2007 was one tragic example of a system failing violently.[78] Similar incidences may be addressed by equipping the system with predefined measures to limit the effects of the system's actions, restrict the system's following actions, or assist and/or compensate those toward which the deploying force has a legal or moral obligation. Examples of remedies include hard-coded capacities to isolate the fire control unit from the remainder of the platform; remotely or manually shut-down; self-shut down upon recognition of a known fault or damage; terminate any action against

protected symbols; and deploy autonomous rescue devices to assist those stranded at sea or in arid areas by the faulty operation of the system.

Not a Checklist but a Guide to Action and Reflection

Obviously, having some moral maxims written on a golden plate does not ensure that AWS are used ethically and should not be seen as a pretext to evade the assessment of possible misuses. We acknowledge that principles, while expressing universally valid values and needs, cannot specify the effective modalities of their own implementation,[79] which depends on contextual considerations that may significantly differ from one warfare scenario to another. That is why our framework is inherently open to further articulation as it is intentionally designed to offer necessary but not sufficient criteria for promoting the most ethically conceivable use of military AI and assisting the continuous betterment of operational protocols and training materials to be developed by practitioners.

A great deal of empirical study, training, and policy making is required to make these 11 principles concretely relevant and effective: the armed forces and all the governmental institutions involved in the adoption of AWS must define rules, standards, and practices to ensure that the deployment and use of AI-based technologies is consistent with the principles; analogously, the technological industry must develop specific capabilities, trust marks, and assessment protocols to keep up to the ethical expectations set by them.

Our suggestion to link the development of ethical standards for military AI to the expertise acquired by private firms challenges fatalist beliefs in the impossibility of maintaining human control. The persistent debate on the international ban of AWS, instead of promoting effective regulations, has encouraged actors toward absolving themselves of moral responsibilities associated with design, development, manufacturing, and maintenance. In turn, core principles are meant to highlight these responsibilities beyond the boundaries of a debate that has remained excessively focused on abstract hypotheses instead of working concretely to make technologies ethically compliant.

The specific nature of autonomous technologies demands that human responsibilities are emphasized, not discounted. Since even the most autonomous weapons have some degree of human interaction in their lifecycle, any AI-enabled military system makes directly or indirectly accountable a large number of people for which any appropriate set of ethical principles should account.[80] A major risk, in absence of clear ethical principles, is that the classical *problem of many hands*, where it is difficult to allocate responsibility among several actors, would become

a *problem of no hands*, an absurd situation in which nobody can be held accountable for the bad design of a weapon due to the lack of explicit ethical criteria.[81]

To address this risk, discussions concerning the ethical application of AI should *be informed by* ethical principles and *contribute to* further *articulating* the implementation of said principles in specific scenarios. These discussions should include not only the military procurer and end-user but also all individuals, from programmers and software engineers to training officers and quality assurance personnel who deal with AI technology. Failing to look beyond the military commander neglects the causal contributions of these actors during the development and use of AWS.

Focusing on small actors does not mean overlooking the role of corporations, states, and intergovernmental agencies, which hold greater capability to effect change, thus, play a prime ethical role. But, in the AI age, the capability to weaponize autonomy depends, for better or worse, on those behind every design input and keystroke. To recognize their role as moral agents, we must promote a professional ethic that includes cautious consideration of the risks associated with all technological media as tools potentially serviceable for military goals and reaches all the way down the command chain to the level of the individual decision makers, irrespective of rank and titles.

This idea is epitomized by the concept of a MinAI system,[82] an AWS architecture that fulfills all the minimal ethical requirements specified by our principles through its capability to tempestively restrict or abort certain offensive actions—for examples, cancelling a human-ordered attack when persons *hors de combat* (surrendered units), policy-identified entities (diplomatic bodies), or an internationally protected symbol (Red Cross)—are detected in the target area.

Compared to a fully-fledged ethical AWS that actively selects among the targets to be acquired, a MinAI system is more likely to meet the national requirements in all Western nations and stands a strong chance to be accepted by other nations (as per principle II), enhancing interoperability when deployed simultaneously by multiple allied forces (as per principle III). Incorporating a system for tracking reliability and human input, as required under principles IV (explainability) and V (human control), a MinAI system would be involved only in justified uses. It would be inherently ethical by design (principle VII) as it would satisfy the requirements set by the three main normative theories of ethics: its effects are preferable to the effects produced by a weapon without AI in an analogous situation (in accord with consequentialist ethics); its design is respectful of certain civil rights and values (as per principle I), unlike an equivalent weapon without AI (in accord with deontology ethics); and it allows personnel to flourish (as per principle IX), as they can fulfill the ethical and legal requirements governing humani-

tarian protection (in accord with virtue ethics). Recent technological advances make the requirements necessary to meet principles VI, X, and XI achievable. Their implementation would constitute a humanitarian enhancement and an illustration of the value of the principles proposed here.

To clarify, we are not claiming that all the possible uses of a MinAI system are unconditionally ethical, as morality is not achieved by mere formal compliance with principles: designing and developing AWS that comply with the MinAI framework does not guarantee that applying lethal force through these AWS would always be the most ethical option.

But this does not make the 11 principles less valid as regulative ideals, i.e., indications of goals and desiderata: these principles would be ethically compelling guides to action and reflection even if the requirements of the MinAI framework were contingently unrealizable due to technological, strategic, or administrative unreadiness. Against the categorical conflation promoted by the ban, we must distinguish the inherent ethical validity of the principles regulating AWS (which draw their normative force from fundamental shared values and societal expectations) and the efforts required to concretely tailor these principles to real-world scenarios (which depend on the capability of technological industry and governmental institutions to fulfill all the relevant practical conditions), remembering that the latter are subordinated to achieving the former; thus, the former are not less valuable when the latter are imperfect or insufficient.

Limitations and Concerns

Our principles provide a framework from which to develop more formal standards and specific codes of practice that are relevant from both a technical and a policy perspective. However, any similar proposal comes with three potential risks: that the principles are too vague to provide a univocal guide to action; that, when applied to concrete scenarios, some latent contradictions between principles might undermine their efficacy; and that the principles, despite being stated clearly, could not be enforced effectively.

The first risk arises when an abstract principle (e.g., "fairness," "inclusiveness") allows different chains of inferences, creating uncertainty about its most appropriate interpretation. Our principles are general enough to apply to a range of relevant scenarios; hence, to guide practical action univocally, they must be complemented by scenario-specific considerations and guidelines, drawing on past actions and outcomes.[83]

The second risk concerns the latent friction between different principles, which may ignite when the principles are concretely applied to real-world situations.[84] For example, "allowing military personnel to flourish" might occasionally compete

with the indication that AI must "be sensitive to the environment" in that the actions conducive to human flourishing may call for the instrumental use of environmental resources. Complex moral trade-offs may be involved in similar cases, but prioritizing among values[85] is possible by referring to broader doctrines, for instance, appealing to just war theory and the Law of Armed Conflict when the use of military AI must be "justified and transparent."[86]

The third risk is the one that motivates the ban on AWS, based on the concern that ethical constraints on military AI would hardly be backed by enforcement, oversight, or serious consequences for deviation.[87] Similar worries (inefficacy, lack of transparency, etc.) are neither novel nor limited to military technology. It is true that it is often unclear what authority is to legitimately enforce the principles, which becomes open then to claims of "ethics washing" where results are not delivered. The solution we suggest is that a group of independent experts must be responsible: oversight bodies typically conduct their monitoring operations in classified contexts, releasing only heavily redacted reports for public consumption, but their audits can be effective if wisely structured and communicated. The collapse of the Google military AI ethics board, with the resulting media coverage and stock-market fluctuations, indicates that expert groups can have a significant impact on global giants, if only through forcing public backlash in the worst cases.

Despite these criticisms, the continued development of ethical regulations for AI is receiving the political support of those at high levels in technology and government spaces. Now it is time that the armed forces and the firms working with them to develop military applications of AI do the same. Our principles provide a high-level framework and shared language through which soldiers, developers, and other stakeholders can profitably discuss the ethical and legal concerns associated with the legitimate, controlled militarization of AI.

Building ethically just AI systems will certainly require more than moral exhortations and a strong personal ethic[88]; to be effective, these principles need to inform all the stages of the R&D process, starting from conceptual design, so that values are directly embedded into the ideation of military technology and percolate through all use-case scenarios.[89] ✪

Dr. Jai C. Galliott

Dr. Galliott is Socio-Technical Futures Analysis Discipline Leader with the Defence Science and Technology Group of the Australian Department of Defence. He is also Honorary Associate Professor at the University of Wollongong. His recent books include *Lethal Autonomous Weapons* (Oxford University Press); Military *Robots: Mapping the Moral Landscape* (Routledge); *Ethics and the Future of Spying: Technology, National Security and Intelligence Collection* (Routledge); *Super Soldiers: The Ethical, Legal and Social Implications* (Routledge); *Commercial Space Exploration: Ethics Policy and Governance* (Routledge); and *Force Short of War in Modern Conflict: Jus ad Vim* (Edinburgh University Press). He is further an Associate at the Centre for International Studies at the University of Oxford.

Dr. Massimiliano L. Cappuccio

Dr. Cappuccio is Senior Researcher and Deputy Director of the Values in Defense and Security Technology Group at the University of New South Wales. As a cognitive philosopher and a technology ethicist, his work focuses on theoretical and applied issues in human performance (skill acquisition and skill disruption), human-robot interaction (social robotics), philosophy and theory of artificial intelligence (the frame problem), from a perspective that integrates embodied, enactive, and extended approaches to cognition. He conducts an intense activity as organizer of academic events, including interdisciplinary workshops (TEMPER) and international conferences (JSSR).

Dr. Austin Wyatt

Dr. Wyatt is a Socio-Technical Futures Analyst with the Defence Science and Technology Group of the Australian Department of Defence. His research focuses on autonomous weapons. His latest published research explores the role of middle powers in arms proliferation, the use of autonomous systems by law enforcement agencies, and the impact of lethal autonomous weapons on strategic stability.

Notes

1. Austin Wyatt and Jai Galliott, "Closing the Capability Gap: Military Modernization During the Dawn of Autonomous Weapon Systems," *Asian Security* 16, no. 1 (2020).

2. Peter Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making," *International review of the Red Cross* 94, no. 886 (2012). "Ban Killer Robots before They Become Weapons of Mass Destruction," *Scientific American*, 7 August 2015.

3. Tim Simonite, "Sorry, Banning "Killer Robots" Just Isn't Practical," *Wired*, 22 August 2017 2017. Matthijs M Maas, "Innovation-Proof Global Governance for Military Artificial Intelligence?: How I Learned to Stop Worrying, and Love the Bot," *Journal of International Humanitarian Legal Studies* 10, no. 1 (2019). Didier Danet, "Do Not Ban "Killer Robots"!" (paper presented at the 2017 International Conference on Military Technologies (ICMT), Brno, Czech Republic, 2017).

4. C. Todd Lopez, "Dod Adopts 5 Principles of Artificial Intelligence Ethics," news release, 25 February 2020, 2020, https://www.defense.gov/News/News-Stories/Article/Article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/.

5. A. Wyatt, "Charting Great Power Progress toward a Lethal Autonomous Weapon System Demonstration Point," *Defence Studies* 20, no. 1 (2020).

6. Kelsey Piper, "Exclusive: Google Cancels Ai Ethics Board in Response to Outcry," *Vox*, 4 April 2019 2019.

7. Sundar Pichai to The Keyword, 7 June 2018, 2018, https://www.blog.google/technology/ai/ai-principles/.

8. Joshua Brustein and Mark Bergen, "Google Wants to Do Business with the Military—Many of Its Employees Don't," *Bloomberg Businessweek*, 25 November 2019 2019.

9. J. Galliott, *Military Robots: Mapping the Moral Landscape* (Surrey, UK: Ashgate, 2015); Carolyn McKay, "Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, Ai and Judicial Decision-Making," *Current Issues in Criminal Justice* 32, no. 1 (2020); Katie Miller, "A Matter of Perspective: Discrimination, Bias, and Inequality in Ai," in *Legal Regulations, Implications, and Issues Surrounding Digital Data*, ed. Margaret Jackson and Marita Shelly (Hershey, Pennsylvania: IGI Global, 2020).

10. Michael C Horowitz, "Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systmes," *Temple International & Comparative Law Journal* 30, no. 1 (2016).no. 1 (2016

11. Loren Thompson, "Five Reasons Why Silicon Valley Won't Partner with the Pentagon," *Forbes*, 27 August 2015 2015.

12. Danet, "Do Not Ban "Killer Robots"!."; Samuel Gibbs, "Elon Musk Leads 116 Experts Calling for Outright Ban of Killer Robots," *The Guardian*, 21 August 2017 2017; Open Robo-Ethics Institute, "The Ethics and Governance of Lethal Autonomous Weapons Systems. An International Public Opinion Poll," (Open Robo-Ethics Institute, 2015).

13. Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making."; "Ban Killer Robots before They Become Weapons of Mass Destruction."; Noel E Sharkey, "The Evitability of Autonomous Robot Warfare," *International Review of the Red Cross* 94, no. 886 (2012); Toby Walsh, *2062: The World That Ai Made* (Carlton: La Trobe University Press, 2018); "Killer Robots: Technological, Legal, and Ethical Challenges," in *Annual Meeting American Association for the Advancement of Science* (Washington D.C.2019).

14. Jai Galliott, "The Limits of Robotic Solutions to Human Challenges in the Land Domain," *Defence Studies* 17, no. 4 (2017); Robert James Sparrow, "War without Virtue?," in *Killing by Remote Control: The Ethics of Unmanned Military*, ed. Bradley J. Strawser (New York: Oxford University Press, 2013); Shannon Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character," *Philosophy & Technology* 28, no. 1 (2015).

15. Daniele Amoroso et al., "Autonomy in Weapon Systems: The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy," in *Publication Series on Democracy* (Heinrich Böll Foundation, 2018).

16. Sharkey, "The Evitability of Autonomous Robot Warfare."

17. Forrest E. Morgan et al., "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World," (Santa Monica, California: RAND Corporation, 2020).

18. Institute, "The Ethics and Governance of Lethal Autonomous Weapons Systems. An International Public Opinion Poll."

19. Amoroso et al., "Autonomy in Weapon Systems: The Military Application of Artificial Intelligence as a Litmus Test for Germany's New Foreign and Security Policy."; Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making."; Sharkey, "The Evitability of Autonomous Robot Warfare."; Walsh, "Killer Robots: Technological, Legal, and Ethical Challenges."

20. Office of the Assistant Secretary of Defense for Acquisition, "Unmanned Systems Integrated Roadmap 2017-2042," ed. Department of Defense (Department of Defense, 2018); Morgan et al., "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World."

21. Gibbs, "Elon Musk Leads 116 Experts Calling for Outright Ban of Killer Robots."

22. Robert Sparrow, "Killer Robots," *Journal of applied philosophy* 24, no. 1 (2007); Gugleilmo Tamburrini, "On Banning Autonomous Weapons Systems: From Deontological to Wide Consequentialist Reasons," in *Autonomous Weapons Systems: Law, Ethics, Policy*, ed. Nehal C. Bhuta, et al. (Cambridge, United Kingdom: Cambridge University Press, 2016).

23. Ronald C Arkin, "The Case for Ethical Autonomy in Unmanned Systems," *Journal of Military Ethics* 9, no. 4 (2010); Galliott, *Military Robots: Mapping the Moral Landscape*; Jason

Scholz, "Ethical Weapons: A Case for Ai in Weapons," in *Moral Responsibility in the Twenty-First Century Just War Theory and the Ethical Challenges of Autonomous Weapons Systems*, ed. Steven C. Roach and Amy E. Eckert (Albany: State University of New York Press, 2020); Jason Scholz and Jai Galliott, "The Case for Ethical Ai in the Military," in *The Oxford Handbook of Ethics of Ai*, ed. Markus D. Dubber, Frank Pasquale, and Sunit Das (New York: Oxford University Press, 2020).

24. Danet, "Do Not Ban "Killer Robots"!."; Marco Sassóli, "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified," *International Law Studies* 90 (2014).

25. MacIntosh, "Fire and Forget: A Moral Defense of the Use of Autonomous Weapons Systems in War and Peace."

26. Vincent C. Müller, "Autonomous Killer Robots Are Probably Good News," in *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*, ed. Ezio Di Nucci and Filippo Santoni de Sio (London: Routledge, 2016).

27. Daniel Trusilo and Thomas Burri, "The Ethical Assessment of Autonomous Systems in Practice," *J* 4, no. 4 (2021).

28. Elvira Rosert and Frank Sauer, "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies," *Contemporary Security Policy* 42, no. 1 (2021).

29. Maas, "Innovation-Proof Global Governance for Military Artificial Intelligence?: How I Learned to Stop Worrying, and Love the Bot."

30. Rosert and Sauer, "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies."

31. Austin Wyatt, "So Just What Is a Killer Robot?: Detailing the Ongoing Debate around Defining Lethal Autonomous Weapon Systems," *Wild Blue Yonder* (2020).

32. A. Wyatt, *The Disruptive Impact of Lethal Autonomous Weapons Systems Diffusion: Modern Melians and the Dawn of Robotic Warriors* (London: Routledge, 2021).

33. Rosert and Sauer, "How (Not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies."

34. Müller, "Autonomous Killer Robots Are Probably Good News."

35. Trusilo and Burri, "The Ethical Assessment of Autonomous Systems in Practice."

36. Acquisition, "Unmanned Systems Integrated Roadmap 2017-2042."

37. Morgan et al., "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World."

38. Maas, "Innovation-Proof Global Governance for Military Artificial Intelligence?: How I Learned to Stop Worrying, and Love the Bot."

39. Tom L Beauchamp, "Principlism and Its Alleged Competitors," *Kennedy Institute of Ethics Journal* 5, no. 3 (1995); Tom Sorell, "The Limits of Principlism and Recourse to Theory: The Example of Telecare," *Ethical theory and moral practice* 14, no. 4 (2011).

40. Jai Galliott and Jason Scholz, "Artificial Intelligence in Weapons: The Moral Imperative for Minimally-Just Autonomy," *US Air Force Journal of Indo-Pacific Affairs* 1, no. 2 (2019).

41. Pichai Ai at Google: Our Principles.

42. Ibid.

43. J. Whittlestone et al., "The Role and Limits of Principles in Ai Ethics: Towards a Focus on the Tensions," in *Conference on Artificial Intelligence, Ethics and Society* (Hawaii2019).

44. Ariel Conn, "An Open Letter to the United Nations Convention on Certain Conventional Weapons," news release, 20 August 2017, 2017, https://futureoflife.org/2017/08/20/autonomous-weapons-open-letter-2017/.

45. ACM US Public Policy Council, "Statement on Algorithmic Transparency and Accountability," news release, 12 January 2017, 2017, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

46. Japanese Society for Artificial Intelligence, "The Japanese Society for Artificial Intelligence Ethical Guidelines," http://ai-elsi.org/archives/514.

47. Université de Montréal, «Montréal Declaration on Responsible Ai,» https://www.montrealdeclarationresponsibleai.com/the-declaration.

48. R. Chatila et al., "The Ieee Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems," *IEEE Robotics and Automation Magazine* 24, no. 1 (2017).

49. Partnerships on AI, "Our Tenets," https://partnershiponai.org/about/.

50. Select Committee on Artificial Intelligence, "Ai in the Uk: Ready, Willing and Able?," (2018).

51. Microsoft, «Microsoft Ai Principles,» https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6.

52. European Commission, "Artificial Intelligence: Commission Takes Forward Its Work on Ethics Guidelines," news release, 8 April 2019, 2019, https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1893.

53. Conn, "An Open Letter to the United Nations Convention on Certain Conventional Weapons."

54. C. Porro, "Ai for Good: Principles I Believe In," Salesforce, https://www.salesforce.org/blog/ai-good-principles-believe/.

55. Jayshree Pandya, "The Weaponization of Artificial Intelligence," *Forbes*, 14 January 2019 2019.

56. Galliott, *Military Robots: Mapping the Moral Landscape*.

57. Ibid.

58. A. Wyatt and J. Galliott, "The Revolution of Autonomous Systems and Its Implications for the Arms Trade," in *Research Handbook on the Arms Trade*, ed. Andrew T. H. Tan (Cheltenham: Edward Elgar, 2020).

59. DST, "The Future of Unmanned Operations Demonstrated at Autonomous Warrior 2018," news release, 19 November 2018, 2018, https://www.dst.defence.gov.au/news/2018/11/19/future-unmanned-operations-demonstrated-autonomous-warrior-2018.

60. Galliott, *Military Robots: Mapping the Moral Landscape*.

61. M. Walzer, *Just and Unjust Wars* (New York: Basic Books, 1987).

62. Jai Galliott, ed. *Force Short of War in Modern Conflict: Jus Ad Vim* (Edinburgh: Edinburgh University Press, 2019).

63. Galliott, *Military Robots: Mapping the Moral Landscape*.

64. Noel Sharkey, "Staying in the Loop: Human Supervisory Control of Weapons," in *Autonomous Weapons Systems: Law, Ethics, Policy*, ed. Nehal C. Bhuta, et al. (Cambridge, United Kingdom: Cambridge University Press, 2016).

65. Galliott, *Military Robots: Mapping the Moral Landscape*.

66. Tim McFarland and Jai Galliott, "Understanding Ai & Autonomy: Problematizing the Meaningful Human Control Argument against Killer Robots," in *Re-Examining the Law and*

Ethics of Lethal Autonomous Weapons, ed. Jai Galliott, J. D. Ohlin, and Duncan MacIntosh (New York: Oxford University Press, 2021).

67. Batya Friedman and David G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination* (New York: MIT Press, 2019).

68. Ibid.

69. Miller, "A Matter of Perspective: Discrimination, Bias, and Inequality in Ai."Margaret</author><author>Shelly, Marita</author></secondary-authors></contributors><titles><title>A matter of perspective: Discrimination, bias, and inequality in ai</title><secondary-title>Legal Regulations, Implications, and Issues Surrounding Digital Data</secondary-title></titles><pages>182-202</pages><dates><year>2020</year></dates><pub-location>Hershey, Pennsylvania</pub-location><publisher>IGI Global</publisher><urls></urls></record></Cite></EndNote>

70. James Zou and Londa Schiebinger to Nature Comment, 18 July 2018, 2018, https://www.nature.com/articles/d41586-018-05707-8.

71. Sheilla Njoto, "Gendered Bots? Bias in the Use of Artificial Intelligence in Recruitment," in *Policy Lab Research Paper* (Melbourne: University of Melbourne, 2020).

72. McKay, "Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, Ai and Judicial Decision-Making."

73. Galliott, "The Limits of Robotic Solutions to Human Challenges in the Land Domain."; Vallor, "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character."

74. Jai Galliott, "Defending Australia in the Digital Age: Toward Full Spectrum Defence," *Defence Studies* 16, no. 2 (2016); "The Unabomber on Robots: The Need for a Philosophy of Technology Geared toward Human Ends," in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, ed. Patrick Lin, Keith Abney, and Ryan Jenkins (New York: Oxford University Press, 2017); "The Soldier's Tolerance for Autonomous Systems," *Paladyn, Journal of Behavioral Robotics* 9, no. 1 (2018).

75. "The Unabomber on Robots: The Need for a Philosophy of Technology Geared toward Human Ends."; "The Soldier's Tolerance for Autonomous Systems."

76. Galliott, *Military Robots: Mapping the Moral Landscape*.

77. Scientific and Cultural Organisation United Nations Educational, "Emblems for the Protection of Cultural Heritage in Times of Armed Conflicts," www.unesco.org/new/en/culture/themes/armed-conflict-and-heritage/convention-and-protocols/blue-shield-emblem/.

78. Noah Shachtman, "Robot Cannon Kille 9, Wounds 14," *Wired*, 10 October 2007. On 2 October 2007, an Oerlikon 35-mm twin-cannon towed antiaircraft gun made malfunctioned, killing nine South African National Defence Force (SANDF) soldiers and injuring 14 more during a training exercise at the SANDF Battle School at Lohatla. A line of eight cannons were engaging a tank hulk in manual ground fire with the guns at low elevation and the maximum traverse of the barrels secured by safety poles and tethers. The rightmost gun jammed during firing, requiring technicians to make repairs. Shortly after the gun was cleared to fire again, the gun malfunctioned, entered automatic mode, broke through the traversal-restriction safety mechanisms. and began firing, striking the other guns along the firing line. Initial reports suggested that the malfunction occurred when the gun underwent an unexplained hang fire of the explosive 35-mm ammunition in the magazines, causing the turret to swing uncontrolled through 360 degrees, firing wildly until it exhausted its remaining ammunition. A statement from South African Defense

Minister, Mosiuoa Lekota, however, stated that the gun had inexplicably traversed 90 degrees to the left, breaking through the safety mechanisms, and fired only a 1⁄8-second-long burst, striking all of the soldiers located on the right-hand side of their guns. An accident report published by the SANDF in January 2008 blamed "undetected mechanical failure—which the manufacturers of an anti-aircraft gun allegedly kept secret." The report says the gun malfunctioned because a spring pin, which is the size of a matchstick, sheared. Other sources blamed poor training and safety procedures in the SANDF. *See* Graeme Hosken, "Army blames gun's maker for Lohatla," *Pretoria News*, 26 January 2008, https://web.archive.org/.

79. Whittlestone et al., "The Role and Limits of Principles in Ai Ethics: Towards a Focus on the Tensions."

80. Morgan et al., "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World."

81. Galliott, *Military Robots: Mapping the Moral Landscape*.

82. Scholz, "Ethical Weapons: A Case for Ai in Weapons."; Scholz and Galliott, "The Case for Ethical Ai in the Military."

83. W. Nicholson Price, "Regulating Black-Box Medicine," *Michigan law review* 116, no. 3 (2017).

84. Beauchamp, "Principlism and Its Alleged Competitors."

85. Ibid.

86. Galliott, *Force Short of War in Modern Conflict: Jus Ad Vim*; Walzer, *Just and Unjust Wars*.

87. M. Whittaker et al., "Ai Now Report 2018," (New York: AI Now Institute, 2018).

88. Whittlestone et al., "The Role and Limits of Principles in Ai Ethics: Towards a Focus on the Tensions."

89. Friedman and Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*.

Disclaimers

The views and opinions expressed or implied in *JIPA* are those of the authors and should not be construed as carrying the official sanction of the Department of Defense, Department of the Air Force, Air Education and Training Command, Air University, or other agencies or departments of the US government or their international equivalents

This article has been cleared for public release by the Australian Department of Defence. The views contained within are those of the authors and do not necessarily represent those of any other party.