CHAPTER 36

THE CASE FOR ETHICAL AI IN THE MILITARY

JASON SCHOLZ AND JAI GALLIOTT

Introduction

Significant recent progress in AI is positively impacting everyday tasks, as well as science, medicine, agriculture, security, finance, law, games, and even creative artistic expression. Nevertheless, some contend that, on ethical grounds, military operations should be immune from the progress of automation and artificial intelligence evident in other areas of society. As an example, Human Rights Watch have stated that:

> Killer robots—fully autonomous weapons that could select and engage targets without human intervention—could be developed within 20 to 30 years... Human Rights Watch and Harvard Law School's International Human Rights Clinic (IHRC) believe that such revolutionary weapons would not be consistent with international humanitarian law and would increase the risk of death or injury to civilians during armed conflict.... The primary concern of Human Rights Watch and IHRC is the impact fully autonomous weapons would have on the protection of civilians during times of war.[1]

The Campaign to Stop Killer Robots, operated by a consortium of nongovernment interest groups, echoes this sentiment, with over 1,000 experts in artificial intelligence, as well as science and technology luminaries such as Stephen Hawking, Elon Musk, Steve Wozniak, Noam Chomsky, Skype co-founder Jaan Tallinn, and Google DeepMind co-founder Demis Hassabis, expressing the problem on their website:

> Allowing life or death decisions to be made by machines crosses a fundamental moral line. Autonomous robots would lack human judgment and the ability to

[1] International Human Rights Clinic, "Losing Humanity: The Case against Killer Robots," Harvard Law School (2012), https://www.hrw.org/sites/default/files/reports/arms1112ForUpload00.pdf.

> understand context. These qualities are necessary to make complex ethical choices on a dynamic battlefield, to distinguish adequately between soldiers and civilians, and to evaluate the proportionality of an attack. As a result, fully autonomous weapons would not meet the requirements of the laws of war. Replacing human troops with machines could make the decision to go to war easier, which would shift the burden of armed conflict further onto civilians. The use of fully autonomous weapons would create an accountability gap as there is no clarity on who would be legally responsible for a robot's actions: the commander, programmer, manufacturer, or robot itself? Without accountability, these parties would have less incentive to ensure robots did not endanger civilians and victims would be left unsatisfied that someone was punished for the harm they experienced.[2]

While we acknowledge some of these concerns, the underlying arguments typically admit no shades of grey, with many based on mistaken assumptions about the role of human agents in the development of these systems and the relevant systems of control. And yet, with such bold arguments from these anti–artificial intelligence luminaries, how can those interested in more nuanced argument begin to rebalance the relevant debate? The anti-AI rhetoric has been permitted to dominate the dialogue on autonomous weapon systems because said debate initially proceeded quite cautiously on the part of the states with responsibility for steering the discussion, on the basis that few understood what it was some were seeking to outlaw with a preemptive ban, but allowing certain advocate groups to sway the debate in the vacuum of informed opinion has given rise to a debate that has ever since been very heavily one-sided.

Meanwhile, with fears about nonexistent sentient robots stalling debate and halting technological progress, one can see in the news that the world faces pressing ethical and humanitarian problems in the use of existing weapons. A gun stolen from a police officer and used to kill, guns used for mass shootings, vehicles used to mow down pedestrians, a bombing of a religious site, a guided-bomb strike on a train bridge as an unexpected passenger train passes over it, a missile strike on a Red Cross facility, and so on. Some of the latter might be prevented by using AI in weapons and in autonomous systems, more generally. It does not seem unreasonable to question why weapons with advanced symbol recognition, for instance, could not be embedded in autonomous systems to identify a symbol of the Red Cross and abort an ordered strike.[3] Similarly, the location of protected sites of religious significance, schools, or hospitals might be programmed into weapons to constrain their actions, or guns prevented from firing by an unauthorized user pointing it at a human. And it does not seem unreasonable to question why this cannot be ensconced in international weapons review standards. We

[2] The Campaign to Stop Killer Robots, "The Solution" (2018), https://www.stopkillerrobots.org/the-solution/.

[3] Indeed, we have introduced the importance of discussing these questions in brief elsewhere. See Jason Scholz and Jai Galliott, "Artificial Intelligence in Weapons: The Moral Imperative for Minimally-Just Autonomy," *US Air Force Journal of Indo-Pacific Affairs* 1 (2018): 57–67. We will also be expanding on the technical feasibility of MinAI, including providing a deployable formal model in a forthcoming book, *Ceding Humanity* (SUNY Press).

seek to correct the lopsided debate and address the concerns of certain advocate groups with a case for a *minimalist* version of *Ethical AI*, explaining why a blanket prohibition on AI in weapons is a bad idea, and why "life" decisions could, and should, at times, be made by machines.

As noted by Lambert and Scholz,[4] automobiles rival wars as a contributor to human death and yet the automobile industry is one of the leaders in integrating automated decision makers into vehicles. Much of the manufacturer's motivation is to make automobiles safer. We hold the same motivation and advocate a similar application in a military context. A simple illustration along the aforementioned lines serves to illustrate. Consider the capability of a weapon to recognize the unexpected presence of an international protection symbol—perhaps a Red Cross, Red Crescent, or Red Crystal—in a defined target area and abort an otherwise unrestrained human-ordered attack. Given the significant advances in visual machine learning over the last decade, such recognition systems are technically feasible. So, inspired by vehicle automation, an Ethical AI system for our purpose is a weapon with inbuilt safety enhancements enabled by the application of artificial intelligence. We further develop this safety argument for weapons, by adapting the guidelines for ethics in autonomous vehicles developed in Germany, but first wish to make the case for Ethical AI in the context of other options, including the impracticability of regulation.

Why Banning Weaponized AI Is a Bad Idea

Autonomous weapons—the primary systems enabled by artificial intelligence—can be serious and dangerous tools in the wrong hands. There is no doubt about this fact. As the above-mentioned tech entrepreneurs and other signatories to a recent open letter to the United Nations have put it, autonomous weapons "can be weapons of terror, weapons that despots and terrorists use against innocent populations, and weapons [that can be] hacked to behave in undesirable ways."[5] But this does not mean that the United Nations ought to proceed immediately to the implementation of a preventive ban on the further development of weaponized artificial intelligence, as the signatories of the open letter urge. For one thing, it sometimes takes dangerous tools to achieve worthy ends.

This is most obvious in the case of humanitarian interventions. Think of the Rwandan genocide, where the world simply stood by and did nothing. Had autonomous weapons capable of discrimination between the relevant fighters been available in 1994, developed

[4] Dale Lambert and Jason Scholz, "A Dialectic for Network Centric Warfare," Proceedings of the 10th International Command and Control Research and Technology Symposium (ICCRTS), MacLean, VA, June 13–16, 2005.

[5] The Future of Life Institute, "An Open Letter to the United Nations Convention on Certain Conventional Weapons" (2017), https://Futureoflife.Org/Autonomous-Weapons-Open-Letter-2017/.

states would likely have been less averse to engagement and may not have looked the other way. It seems plausible that if the costs of humanitarian interventions were purely monetary, that is, if we were removed the sometimes controversial nature of weapons deployment on foreign soil and the concerns that some hold regarding casualty aversion, then it would be easier to gain widespread support for what are otherwise might otherwise be morally sanctioned interventions.[6]

To make this point more generally, it should be acknowledged that AI technology is tremendously beneficial, and it already permeates our lives in ways that people often do not notice and often are not well placed or able to comprehend fully. Given its pervasive presence and the virtual impossibility of constraining a software-underwritten technology that is already in the public domain, it is shortsighted or perhaps even naive to think that the artificial intelligence technology's abuse can be prevented if only the further development of autonomous weapons is halted.

If a ban were to be implemented, the likely consequence would be the development of artificial intelligence–enabled weapons by malicious nonstate and state-based actors using existing technology. It is worth bearing in mind that most artificial intelligence in weapons is currently deployed by developed states that conduct their military and security engagements broadly in line with international law and public expectations, with said technology therefore accompanied by robust safety mechanisms and deployed in appropriate zones given the known limitations of the technology—for example, out at sea rather than in urban conflict zones. Nefarious actors will have no reason to act in such a constrained fashion, and there exists no effective enforcement regime to hold these actors responsible for violations of international law, meaning any prevailing autonomous future of this kind is likely to be bleak and consist of technology minus existing safeguards. In fact, it may well take the sophisticated and discriminate autonomous-weapons systems that developed military forces around the world are currently in the process of developing or, in some cases, deploying, if we are to effectively counter the much cruder autonomous weapons that would likely be constructed through the reprogramming of seemingly benign AI technology such as the self-driving car and other off-the-shelf technologies if a "preventative" ban were to be implemented. The developed states of world, while together an imperfect moral arbiter, have a moral obligation to develop new technologies partly on the basis that it has the responsibility to its collective population to quell the uprising of this crude technology by those who seek to do harm to the many.[7] This is to say that a consequence of a ban would be to deny the use of AI weapons as a countermeasure against other AI or autonomous weapons.

The world has previously placed prohibitions on the possession and use of certain types of weapons, including chemical, biological, nuclear, and potentially persistent unexploded ordnance such as cluster munitions and landmines. Prohibition of these weapons has not prevented states or nonstate actors from developing them. India, Pakistan, Israel, and North Korea have developed nuclear weapons, and Iran was

[6] Jai Galliott, *Military Robots: Mapping the Moral Landscape* (Farnham, UK: Ashgate, 2015).

[7] *Id.* at 37–64.

actively developing a nuclear weapons program until 2009.[8] Moreover, none of the nations that possess nuclear weapons is a signatory to the Treaty on the Prohibition of Nuclear Weapons.[9] Nevertheless, preventing nations or nonstate actors from acquiring nuclear weapons has been reasonably effective until now, but only because it has been possible to physically control access to the relatively difficult-to-obtain materials required to produce them. In the case of AI and autonomous weapons, it is not the materials that are lacking, but the code. The algorithms needed for autonomous weapons are in many cases the same as those needed for autonomous cars or mobile phone apps, so one faces a dual use definitional problem. It is not possible to identify certain types of code that are militarily useful and ban them. The construction of autonomous weapons once the component technologies—many of which will be in the public domain—become available is only a matter of time, and not only for nation states. This is an area in which those states charged with maintaining international order do not want to find themselves lagging behind.

A blanket prohibition on "AI in weapons" would have further unintended consequences due to its lack of *nuance*. Building on the earlier discussion of the implications of halting the development of AI, there is a distinction to be made in any regulation or policy about those *kinds of AI* that could yield significant humanitarian benefits. This lack of nuance is also evident in the case against chemical weapons. For example, pepper spray or tear gas is a chemical agent banned in warfare under the Chemical Weapons Convention of 1993, making it illegal for use by militaries except in law enforcement. The denial of tear gas to military forces removes a less-than-lethal option from the inventory, which could lead to the unnecessary use of lethal force. Even the responsible development of what are often seen as abhorrent weapons can be defended on the basis that they might prevent the use of more deadly or indiscriminate force. Moving along the spectrum of destructive weapons one finds land mines. The United States, of course, never ratified the Ottawa Treaty but rather chose a technological solution to end the use of persistent land mines—land mines that can be set to self-destruct or deactivate after a predefined time period—making them considerably less problematic when used in clearly demarcated and confined zones such as the Korean Demilitarized Zone (DMZ).[10] In choosing not to ratify the Ottawa Treaty, the United States had identified that what, in one form, can be a crude and indiscriminate weapon can, in the hands of the morally scrupulous, be another weapon that may limit the need for more injurious weapons prevent the use of even more deadly and indiscriminate application force in places like the Korean DMZ, where the alternatives might include options that are not sensitive to discrimination between a child (either now or two decades in the future)

[8] Rod Barton, *The Weapons Detective: The Inside Story of Australia's Top Weapons Inspector* (Melbourne: Black Inc. Agenda, 2015).

[9] Alexander White and Matthew Paterson, "Nuke Kid in Town: How Much Does the Treaty on the Prohibition of Nuclear Weapons Actually Change?," *Pandora's Box* 24 (2017): 141–156.

[10] Lorraine Boissoneault, "The Historic Innovation of Land Mines—And Why We've Struggled to Get Rid of Them," *Smithsonian* (February 24, 2017), https://www.smithsonianmag.com/innovation/historic-innovation-land-minesand-why-weve-struggled-get-rid-them-180962276/.

and a military-aged adult during a period of defined hostility, as in the case of modern land mines.

There is also a need to overcome another common notion behind a ban, that which revolves around an overly optimistic view of technology in that it raises concerns regarding a lack of human control. This is a conception that fails to acknowledge the long causal backstory of institutional arrangements and individual actors who, through thousands of little acts of commission and omission in the process of design, engineering, and development, have brought about, and continue to bring about, the rise of such technologies. As long as the debate about autonomous weapons is framed primarily in terms of UN-level policies, the average citizen, soldier, or programmer must be forgiven for assuming that he or she is absolved of all moral responsibility for the wrongful harm that autonomous weapons risk causing. But this assumption is false, and it might prove disastrous. All individuals who deal with AI technology have to exercise due diligence, and each and every one of us needs to examine carefully how his or her actions and inactions are contributing to the potential dangers of this technology and those in which it may be integrated. This is by no means to say that state and intergovernmental agencies do not have an important role to play as well. Rather, it is to emphasize that if the potential dangers of autonomous weapons are to be mitigated, then an ethic of personal responsibility must be promoted, and it must reach all the way down to the level of the individual decision maker. For a start, it is of the utmost importance that we begin telling a richer and more complex story about the rise of AI weapons—a story that includes the causal contributions of decision makers at all levels. From there, we can see how Ethical AI would serve to enhance accountability. Take one example of Ethical AI, "smart guns" that remain locked unless held by an authorized user via biometric or token technologies to curtail accidental firings and cases of a gun stolen and used immediately to shoot people. Or a similar AI mechanism built into any military weapon, noting that even the most autonomous weapons have some degree of human interaction in their life cycle. These technologies might also record events, including the time and location of every shot fired, providing some accountability.

The point here is that the world has large stockpiles of weapons—bombs, mines, bullets, guns, grenades, mortars, and missiles—that have no inbuilt technical controls related to the conditions under which they are employed. This is perhaps a far more frightening reality of immediate humanitarian concern than any fictional scenario involving "killer robots" or out-of-control artificial intelligence. Munitions developed for use by militaries and the public generally possess no inbuilt safeguards that prevent them from being used by unauthorized persons. We must remember that military forces that cannot afford precision weapons are regularly legally justified in the defense of their state to kill enemy combatants with firearms, bombs, and other sometimes imprecise and indiscriminate weapons. Yet as military technology becomes increasingly capable of yielding more precise outcomes at lower cost and halting an enemy without causing unnecessary suffering or harm to those nearby, this is a situation that moral philosophers and international law might now reconsider, and we think this is best done through the lens of Ethical AI.

The Ethical AI Spectrum

A weapon with Ethical AI takes an attack order as input and makes a decision *not to obey* the order if it assesses the presence of unexpected[11] *protected* object(s). What we mean by protected may include legally identified entities from ICRC marked objects, through to persons hors de combat and policy-identified entities specified in rules of engagement. We recognize that ends of this spectrum range from easy to very difficult technological challenges for AI.[12] What this does mean is that some progress toward Ethical AI can be made immediately, and we have proposed a technical model elsewhere. Clearly, any progress would constitute a humanitarian enhancement.

Lambert[13] termed weapon systems with these ethical improvements "Moral Weapons" and included this to mean "fully integrated human-machine decision making," the option of "allowing the machine to at times override the human," the ability to assess and *decline* targeting requests when rules of engagement violations are deduced, with the decisions to override these weapons logged for subsequent accountability review. We term these Ethical AI rather than Moral AI, to avoid any potential confusion with Moral Responsibility, that is, we do not mean to imply that such weapons possess *moral responsibility*.

We assert that AI in weapons is not likely to be banned regardless of campaign efforts, and we advocate that critics or those who generally reject the concept of autonomous weapons might consider this new concept to further reduce casualties over current weapons and address their central concern about humans losing control over decision-making in warfare.

Let us discern between two ends of a spectrum of ethical capability. A maximally just "ethical machine" (MaxAI) guided by both acceptable and nonacceptable actions has the benefit of ensuring that ethically obligatory lethal action is taken, even when system engineers of a lesser system may not have recognized the need or possibility of the relevant lethal action. However, a maximally just ethical robot requires extensive ethical engineering. Arkin's "ethical governor" represents probably the most advanced prototype effort toward a maximally just system.[14] The ethical governor provides assessment on proposed lethal actions consistent with the laws of war and rules of engagement. The maximally just position is apparent from the explanation of the operation of the

[11] One may argue that adversaries who know this might "game" the weapons by posing under the cover of "protection." If this is known, it is a case for (accountable) human override of the Ethical Weapon, and why we use the term "unexpected." Noting also that besides being an act of perfidy in the case of such use of protected symbols, which has other possible consequences for the perpetrators, it may in fact aid in targeting as these would be anomalies with respect to known Red Cross locations.

[12] Robert Sparrow, "Robots and Respect: Assessing the Case against Autonomous Weapon Systems," *Ethics & International Affairs* 30 (2016): 93–116.

[13] Dale Lambert, "Ubiquitous Command and Control," *Proceedings of the 1999 Information, Decision and Control Conference*, Adelaide, Australia (IEEE, 1999), 35–40.

[14] Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton, FL: CRC Press, 2009).

constraint interpreter, which is a key part of the governor: "The constraint application process is responsible for reasoning about the active ethical constraints and ensuring that the resulting behavior of the robot is ethically permissible."[15] That is, the constraint system, based on complex deontic and predicate logic, evaluates the proposed actions generated by the tactical reasoning engine of the system based on an equally complex data structure. Reasoning about the full scope of what is *ethically permissible under all possible conditions* including general distinction of combatants from noncombatants, proportionality, unnecessary suffering, and rules of engagement, as Arkin describes, is a hard problem.

In contrast, a MinAI "ethical robot," while still a constraint driven system, could operate without an "ethical governor" proper and need only contain an elementary suppressor of human-generated lethal action. Further, as it would activate in accordance with a much narrower set of constraints it may be hard- rather than soft-coded, meaning far less system "interpretation" would be required. MinAI deals with what is *ethically impermissible*. Thus, we assert under *certain specific conditions*, distinction, proportionality, and protected conditions may be assessed, as follows:

- *Distinction of the ethically impermissible* including the avoidance of application of force against "protected" things such as objects and persons marked with the protected symbols of the Red Cross, as well as protected locations, recognizable protected behaviors such as the desire to parlay, basic signs of surrender (including beacons), and potentially those that are hors de combat or are clearly noncombatants; noting of course that AI solutions range here from easy to more difficult—but not impossible—and will continue to improve along with AI technologies.
- *Ethical reduction in proportionality* includes a reduction in the degree of force below the level lawfully authorized if it is determined to be sufficient to meet military necessity.

MinAI then is three things: (1) an ethical control to augment any conventional weapon; (2) a system limited to decision and action on logical negative cases of things that should *not* be attacked; and (3) practically achievable with state of the art AI techniques.

The basic technical concept for a MinAI Ethical Weapon is an augmentation to a standard weapon control system. The weapon seeker, which may be augmented with other sensors, provides input to an ethical and legal perception-action system. This system uses training data, developed, tested, and certified prior to the operation and outputs a decision state to override the target order and generate alternate orders on the control system in the event of a world state which satisfies MinAI conditions. The decision override is intended to divert the weapon to another target, or a preoperation-specified fail-safe location and/or to neutralize or reduce the payload effect accordingly.

[15] Ronald C. Arkin, Patrick Ulam, and Brittany Duncan, *An Ethical Governor for Constraining Lethal Action in an Autonomous System* (Fort Belvoir, VA: Defense Technical Information Center, 1 January 2009), https://doi.org/10.21236/ADA493563.

Noteworthy is that while MinAI will always be more limited in technical nature, it may be more morally desirable in that it will yield outcomes that are as good as or possibly even better than MaxAI in a range of specific circumstances. The former will never take active lethal or nonlethal action to harm protected persons or infrastructure. In contrast, MaxAI involves the codification of normative values into rule sets and the interpretation of a wide range of inputs through the application of complex and potentially imperfect machine logic. This more complex "algorithmic morality," while potentially desirable in some circumstances, involves a greater possibility of actively introducing fatal errors, particularly in terms of managing conflicts between interests.

Cognizant of the foregoing information, our suggestion is that in terms of meeting our fundamental moral obligations to humanity, we are ethically justified to develop MinAI systems. The ethical agency of said system, while embedded in the machine and thus technologically mediated by the design, engineering, and operational environment, is fewer steps removed from human moral agency than in a MaxAI system. We would suggest that MaxAI development is supererogatory in the sense that it may be morally beneficial in particular circumstances, but is not necessarily morally required, and may even be demonstrated to be unethical.

The Technical Feasibility of MinAI

To the distaste of some, it might be argued that the moral desirability of MinAI will decrease in the near future as the AI underpinning MaxAI becomes more robust and we move away from rule-based and basic neural network systems toward artificial general intelligence (AGI), and that resources should therefore be dedicated to the development of maximal "ethical robots." To be clear, there have been a number of algorithm success stories announced in recent years, across all the cognate disciplines. Much attention has been given to the ongoing development of algorithms as a basis for the success of AlphaGo[16] and Libratus.[17] These systems are competing against the best human Go and poker players and winning against those who have made acquiring deep knowledge of these games their life's work. The result of these preliminary successes has been a dramatic increase in media reporting on, and interest in, the potential opportunities and pitfalls associated with the development of AI, not all of which are accurate and some of which has negatively impacted public perception of AI, fueling the kind of dystopian visions advanced by the Campaign to Stop Killer Robots.

The speculation that superintelligence is on the foreseeable horizon, with AGI timelines in the realm of twenty to thirty years, reflects the success stories while omitting

[16] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aga Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, and Yutian Chen, "Mastering the Game of Go without Human Knowledge," *Nature* 550 (2017): 354–359.

[17] Noam Brown and Tuomas Sandholm, "Superhuman AI for Heads-Up No-Limit Poker: Libratus Beats Top Professionals," *Science* 359 (2018): 418–424.

discussion of recent failures in AI. Many of these undoubtedly go unreported for commercial and classification reasons, but Microsoft's Tay AI Bot, a machine learning chatbot that learns from interactions with digital users, is but one example.[18] After a short period of operation, Tay developed an "ego" or "character" that was strongly sexual and racialized, and ultimately had to be withdrawn from service. Facebook had similar problems with its AI message chatbots assuming undesirable characteristics,[19] and a number of autonomous road vehicles have now been involved in motor vehicle accidents where the relevant systems were incapable handling the scenario[20] and quality assurance practices failed to factor for such events.

There are also known and currently irresolvable problems with the complex neural networks on which the successes in AI have mostly been based. These bottom-up systems can learn well in tight domains and easily outperform humans in these scenarios based on data structures and their correlations, but they cannot match the top-down rationalizing power of human beings in more open domains such as road systems and conflict zones. Such systems are risky in these environments because they require strict compliance with laws and regulations, and it would be difficult to question, interpret, explain, supervise, and control them by virtue of the fact that deep learning systems cannot easily track their own "reasoning."[21]

Just as importantly, when more intuitive and therefore less explainable systems come into wide operation, it may not be so easy to revert to earlier stage systems as human operators become reliant on the system to make difficult decisions, with the danger that their own moral decision-making skills may have deteriorated over time.[22,23] In the event of failure, total system collapse could occur with devastating consequences if such systems were committed to mission-critical operation required in armed conflict.

There are, moreover, issues associated with functional complexity and the practical computational limits imposed on mobile systems that need to be capable of independent operation in the event of a communications failure. The computers required for AGI-level systems may not be subject to miniaturization or simply may not be sufficiently powerful or cost-effective for the intended purpose, especially in a military context in

[18] Rafal Rzepka and Kenzi Araki, "The Importance of Contextual Knowledge in Artificial Moral Agents Development," *AAAI Spring Symposium Series*, North America (2018), https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/view/17540/15376.

[19] Erin Griffith and Tom Simonite, "Facebook's Virtual Assistant M Is Dead. So Are Chatbots," *Wired* (January 8, 2018), https://www.wired.com/story/facebooks-virtual-assistant-m-is-dead-so-are-chatbots/.

[20] Francesca Favarò, Sky Eurich, and Nazanin Nader, "Autonomous Vehicles' Disengagements: Trends, Triggers, and Regulatory Limitations," *Accident Analysis & Prevention* 110 (2018): 136–148.

[21] Martin Ciupa, "Is AI in Jeopardy? The Need to Under Promise and Over Deliver—The Case for Really Useful Machine Learning," in *4th Int. Conf. on Computer Science and Information Technology*, ed. Dhinaharan Nagamalai et al. (AIRCC, 2017), 59–70.

[22] Jai Galliott, "The Limits of Robotic Solutions to Human Challenges in the Land Domain," *Defence Studies* 17 (2017): 327–345.

[23] Jai Galliott, "Defending Australia in the Digital Age: Toward Full Spectrum Defence," *Defence Studies* 16 (2016): 157–175.

which autonomous weapons are sometimes considered disposable platforms.[24] The hope for advocates of AGI is that computer-processing power and other system components will continue to become dramatically smaller, cheaper, and powerful, but there is no guarantee that Moore's law, which supports such expectations, will continue to reign true without extensive progress in the field of quantum computing.

MaxAI at this point in time, whether or not AGI should eventuate, appears a distant goal to deliver a potential result that is far from guaranteed. A MinAI system, on the other hand, seeks to ensure that the obvious and uncontroversial benefits of artificial intelligence are harnessed while the associated risks are kept under control by normal military targeting processes. Action needs to be taken now to intercept grandiose visions that may not eventuate and instead deliver a positive result with technology that already exists.

A Code for MinAI

A positive result for MinAI will also require more fine-grained guidance on the system's implementation and application. In 2013 the Human Rights Council of the UN General Assembly made the recommendation that developers of lethal autonomous robots (LARs) of the kind enabled by AI "establish a code or codes of conduct, ethics and/or practice defining responsible behaviour with respect to LARs."[25]

As a starting point, one might look for similar codes in related fields. In July 2016, the German Federal Ministry of Transport and Digital Infrastructure (*Bundesministerium für Verkehr und digitale Infrastruktur*, BMVI) appointed an expert panel of scientists and legal experts to serve as a national ethics committee for autonomous vehicles.[26] A year later they made headlines when they issued "the world's first ethical guidelines for driverless cars."[27] Obviously, automobiles are not designed to be weapons, though their kinetic energy and ubiquity make them at least as deadly in practice, such that their automation raises a number of issues in terms of potential damage to life and property. Many of the normative questions that arise as a result, and the normative frameworks utilized to answer said questions, are similar. As such, it does not seem unreasonable to take the BMVI ethics code as a basis for the development of an analogous code for Ethical AI. This may be further justified after consideration of some of the relevant principles of the Law of Armed Conflict (LOAC):

[24] Ciupa, "AI in Jeopardy."

[25] Christoph Heyns, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, A/HRC/23/47 (Geneva: United Nations General Assembly Human Rights Council, 2013).

[26] Christoph Leutge, "The German Ethics Code for Automated and Connected Driving," *Philosophy and Technology* 30 (2017): 547–558.

[27] David Tuffley, "At Last! The World's First Ethical Guidelines for Driverless Cars," *The Conversation* (September 3, 2017), https://theconversation.com/at-last-the-worlds-first-ethical-guidelines-for-driverless-cars-83227.

Military necessity. For military operations, any use of weapons requires said use to produce military gains that are not otherwise prohibited by international humanitarian law.[28] The BMVI code does not address this issue, since the presumption is that automobiles have a right to be on the road for purposes of transport regardless of what or whom they are transporting, and thus needs to be augmented as part of LOAC.

Distinction. The ability to distinguish between the civilian population and combatants, and between civilian objects and military objectives, and accordingly direct operations only against military objectives.[29] In the case of Ethical Weapons, they might identify protected symbols, noncombatants, surrendering persons, and persons who are hors de combat in order accordingly as (and if) the AI technologies continue to advance. Distinction is not used in the German automobile ethics code, except in the priority for human persons over nonhuman persons (i.e., animals) in the case of an impending accident. This again, is included under LOAC.

Proportionality. In armed conflict, some noncivilian casualties may be justifiable in certain circumstances, as long as they are not excessive in relation to the anticipated military advantage. These are illustrated in the subject of automotive trolley problem studies and are included in the German ethics guidelines.[30] But how much of an obligation do military strategists have to avoid harm to civilian populations? Customary IHL provides further useful protections beyond merely justifying proportionality on the basis of the principle of double effect including: Rule 15 (precautions in attack), Rule 20 (advance warning), and Rule 24 (removal of civilians and civilian objects), which are applied in the following section to Ethical Weapons.

The German ethics code opens with general remarks and a mission statement. We have adapted this as follows.

Ethical MinAI Mission

Important decisions will have to be made concerning the extent to which the use of Ethical AI in weapons is required. States have a record of failing to intervene with new weapons technologies, even when doing so would have been justified. The character of the justification to employ Ethical Weapons could be understood in three ways.

First, states with the capability and capacity to do so may be obliged to deploy Ethical Weapons and hence face blame should they decide otherwise. An argument for these capabilities potentially being obligatory is that ethical weapons improve humanitarian outcomes (reducing accidental deaths, etc.) without impact on military effectiveness

[28] International Committee of the Red Cross, *Declaration Renouncing the Use, in Time of War, of Explosive Projectiles under 400 Grammes Weight*, Saint Petersburg (1868).

[29] International Committee of the Red Cross, Geneva Conventions of 1949 and Additional Protocols, and Their Commentaries.

[30] Jean-Francois Bonnefon, Azim Shariff, and Iyad Rahwan, "The Social Dilemma of Autonomous Vehicles," *Science* 352 (2016): 1573–1576.

and are likely to utilize technologies that are low-cost due to their commercial scale, with further justification explained by Galliott.[31]

Second, the development and deployment of Ethical AI could be supererogatory in the sense that it would be good for a state to intervene with Ethical AI–enabled weapons in particular circumstances, but not ethically required.

Third, such action could be justified but neither obligatory nor supererogatory, such that the use of Ethical AI weapons would be ethically acceptable but likely to yield little benefit over the status quo. We suggest that in all cases where the use of Ethical AI weapons is justified, that is, in the pursuit of just causes, their use is either ethically obligatory or supererogatory, but much hinges on the conditions in which they are used and the way in which they are designed.

At a fundamental level, the deployment decision can be reduced to a few fundamental questions: How much dependence on technologically complex systems—based on artificial intelligence and machine learning—are we willing to accept in order to achieve, in return, more safety for noncombatants, more safety for our military, who, acting on behalf of our society, warrant protection, better compliance with laws of armed conflict, and improved operational efficiency to defeat ever improving adversary capabilities? What precautions are needed to ensure appropriate competency, authority, and responsibility? What technological development guidelines are required to ensure that we do not blur the contours of a human society that places trust in its military commanders and their freedom of action, physical and intellectual integrity, and entitlement to social respect at the heart of its legal regime?

In what follows, we propose fourteen principles to guide the development of the MinAI from concept to technical implementation, as adapted to a military context.

Ethical Guidelines

1. Purpose

The primary purpose of Ethical AI is to improve the safety of protected entities and noncombatants within the Law of Armed Conflict and rules of engagement. A secondary purpose is to increase freedom of maneuver for military commanders, thereby enabling further ethical benefits.

2. Positive Balance of Risks

The objective is to reduce the level of harm within the Laws of Armed Conflict with the ultimate goal of zero unintended noncombatant casualties. The fog of war means that

[31] Galliott, *Military Robots*, ch. 3.

noncombatant casualties will be a reality in twenty-first-century warfare, but to minimize these toward zero should be the ultimate aim, made possible only by increasing the intelligence of weapons, projectiles, and effectors of all kinds. The adoption of Ethical AI is justifiable if it promises to produce a diminution in harm to human and/or political capital in comparison to conventional weapons.

3. Avoidance of Ethical Dilemmas to the Extent Possible

Ethical AI should prevent noncombatant harm within the Laws of Armed Conflict wherever this is practically possible. Further, appropriate reduction in operator involvement might reduce risk of post-traumatic stress disorder. Based on the state of the art, the technology should be designed in such a way that critical situations do not arise in the first place. These include dilemma situations, in which Ethical AI and/or military commanders have to decide which of two "evils" to perform. In this context, the entire spectrum of technological options should be used and continuously evolved; for example, limiting the scope to certain controllable conditions in military environments, allowing the weapon to dynamically and cognitively choose a payload yield reduction below a maximum level authorized, making the payload inert, performing weapon avoidance maneuvers, producing signals or advance warnings for persons at risk, or deferring strike to alternate points of opportunity in time and space. The significant enhancement of noncombatant safety is the objective of development and regulation, starting with design and programming of the Ethical AI such that it tracks in a defensive and anticipatory manner, posing as little risk as possible to vulnerable noncombatants while still achieving its missions.

4. Armed Conflict Shall Be Managed by Mixed Initiative Agreements

A statutorily imposed obligation to use Ethical AI is ethically questionable if it entails submission of *all* military commanders to technological imperatives. That is, there should be a prohibition on degrading humans to *only* being subservient elements in an autonomous network. Dynamic and recorded mixed-initiative agreements between humans and machines shall subsume hierarchical human-only command arrangements for Ethical AI.

5. Primacy of Human Life

In situations that prove to be unavoidable, despite all technological precautions being taken, protection of humans enjoys priority in a balancing of interests compared with damage to animals or property.

6. Military Commanders Decide to Sacrifice Specific Lives

Ethical AI can execute targeting according to processes approved by military commanders in compliance with laws of armed conflict and rules of engagement.

7. Machines Minimize Innocent Casualties

In the event of situations where the death of innocent people is unavoidable, Ethical AI shall seek to minimize casualties among innocent people.

8. Military Commanders, Developers, and Defense Departments Are Accountable for Ethical Weapons

Military commanders throughout the network of command remain accountable for the use of Ethical AI. All Ethical AI systems will log the protocol exchange between Ethical AI and military personnel, as well as critical weapon status and knowledge, to provide accountability and postaction review from the perspectives of accountability of commanders, developers, and defense departments as a whole.

9. Security of Ethical Weapons

Ethical AI is justifiable only to the extent that conceivable attacks, in particular manipulation of the information technologies it relies upon or other innate system weaknesses, do not result in such harm as to undermine confidence in the military or in Ethical AI.

10. Awareness and Recording of Responsibility Transfers

It must be possible to clearly distinguish whether an Ethical AI system is being used, where accountability lies and that it comes with the option of overruling the system. The human-machine interface must be designed such that it is clearly regulated and apparent where authority, competency, and responsibility lies, especially the responsibility for control. The distribution of responsibilities (and thus of accountability), for instance with regard to the time and access arrangements, should be reliably recorded and stored. This applies especially to human-to-technology handover procedures.

11. Human On- and Off-the-Loop

The software and technology associated with Ethical AI must be designed such that the need for an abrupt handover of control to military commanders is minimized. To enable

efficient, reliable, and secure human-machine communication and prevent overload, the systems should adapt to human communicative behavior where possible, rather than requiring humans to enhance their adaptive capabilities. Communication to the human will be appropriately abstracted and sufficiently timely where feasible, noting that human-in-the-loop will give way to human-on-the-loop, and human-off-the-loop relationships for periods of time.

12. Machine Self-Learning Considerations

Learning systems that are self-learning in training, operation, and their connection to scenario databases may be allowed if, and to the extent that, they generate safety gains. Self-learning systems must not be deployed unless they meet the safety requirements for Ethical AI and do not undermine these guidelines.

13. Fail Safe Management

In situations where protected marked objects, or unanticipated noncombatants are present, Ethical AI must autonomously (i.e., without direct human intervention) enter into a "safe condition." Identification of what constitutes safe conditions for weapon disposal and recovery, planning, and handover routines is required prior to Ethical AI use. This may include means under control of the machine to: place the weapon in a location that has minimal human impact; neutralize explosives in the weapon, for example, by use of separated chemical components in warhead design which are diffused to prevent future ignition or exploitation; and reduce weapon kinetic energy and damage.

14. Military Education and Training

The proper use of Ethical AI should form part of military commanders' general education. The proper handling of Ethical AI should be taught in an appropriate manner during training, and teams of commanders and Ethical AI tested for capability certification.

Potential Consequences of MinAI

Concerns may be raised that should MinAI functionality be adopted for use by military forces, the technology may result in negative or positive unintended long-term consequences. If so, what might these be? Conscious of how notoriously difficult it is to predict technology use, this is not an easy question to answer, but we will consider some important cases for further study.

Complacency and Responsibility Transfer

One possible negative affect is related to human complacency. Consider the hypothesis "if MinAI technology works well and is trusted, its operators will become complacent in regard to its use and take less care in the targeting process, leading to more deaths."

In response, such an argument would apply equally to all uses of technology in the targeting process. Clearly however, technology is a critical enabler of intelligence and targeting functions. Complacency then seems to be a matter of adequate discipline, appropriate education, training, and system design.[32]

A less desirable outcome would be for operators to abdicate their responsibilities for targeting. Campaigners have attempted to argue the creation of a "responsibility gap" in autonomous weapons before; might this resurface with the application of a MinAI system? Consider the hypothesis that "if MinAI technology works well and is trusted, that Commanders might just as well authorize weapon release with the highest possible explosive payload to account for the worst case and rely on MinAI to reduce the yield according to whatever situation the system finds to be the case, leading to more deaths."

In response to this argument, we assert that this would be like treating MinAI weapon system as if it were a MaxAI weapon system. We do not advocate MaxAI weapons. A MinAI weapon that can reduce its explosive payload under AI control is not a substitute for target analysis; it is a last line of defense against unintended harm. Further the commander would remain responsible for the result, regardless, under any lawful scheme. Any weapon can be misused. A machine gun can he used by a soldier in combat or deployed by a civilian in a school shooting. Discipline, education, and training remain critical to the responsible use of weapons.

Denying Availability of Surrender Technology to Combatants

If the machine-recognizable surrender system were to be developed, would militaries not want to issue beacons to their soldiers because they fear mass surrender? This could prove to be of positive military and/or moral benefit when viewed objectively. It could be mandated that beacons be offered to all soldiers. Stigma or culture associated with use or underuse would, of course, present some degree of concern.

Conclusion

We have presented a case for autonomy in weapons that could make life-saving decisions in the world today. Minimally Just Ethical AI in weapons should achieve a reduction in accidental strikes on protected persons and objects, reduce unintended strikes

[32] Galliott, "The Limits of Robotic Solutions."

against noncombatants, reduce collateral damage by reducing payload delivery, and save lives of those who have surrendered.

We hope in future that the significant resources spent on reacting to speculative fears of campaigners might one day be spent mitigating the definitive suffering of people caused by weapons which lack minimally just autonomy based on artificial intelligence.

Bibliography

Arkin, Ronald. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, FL: CRC Press, 2009.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathon Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "The Moral Machine Experiment." *Nature* 563 (2018): 59–64.

Galliott, Jai. *Military Robots: Mapping the Moral Landscape*. Farnham, UK: Ashgate, 2015.

Galliott, Jai. "Defending Australia in the Digital Age: Toward Full Spectrum Defence." *Defence Studies* 16 (2016): 157–175.

Galliott, Jai. "The Limits of Robotic Solutions to Human Challenges in the Land Domain." *Defence Studies* 17 (2017): 327–345.

Leben, Derek. *Ethics for Robots: How to Design a Moral Algorithm*. New York: Routledge, 2018.

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. New York: Oxford University Press, 2017.

Lin, Patrick, George Bekey, and Keith Abney. *Autonomous Military Robotics: Risk, Ethics, and Design*. Washington, DC: United States Department of the Navy, 2008.

Scholz, Jason and Jai Galliott. "Artificial Intelligence in Weapons: The Moral Imperative for Minimally-Just Autonomy." *US Air Force Journal of Indo-Pacific Affairs* 1 (2018): 57–67.

Sparrow, Robert. "Building a Better WarBot: Ethical Issues in the Design of Unmanned Systems for Military Application." *Science and Engineering Ethics* 15 (2009): 169–187.