frontiers | Research Topics

Ethical challenges in AI-enhanced military operations

Edited by
George Lucas, Henrik Syse, Kirsi Helkala and Edward Barrett

Published in
Frontiers in Big Data

ISSN 1664-8714
ISBN 978-2-8325-2896-9
DOI 10.3389/978-2-8325-2896-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Ethical challenges in AI-enhanced military operations

Topic editors

George Lucas — United States Naval Academy, United States
Henrik Syse — Peace Research Institute Oslo (PRIO), Norway
Kirsi Helkala — Norwegian Defence University College, Norway
Edward Barrett — United States Naval Academy, United States

Citation

Lucas, G., Syse, H., Helkala, K., Barrett, E., eds. (2023). *Ethical challenges in AI-enhanced military operations*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-2896-9

Table of contents

frontiers | Frontiers in Big Data

TYPE Editorial
PUBLISHED 19 June 2023
DOI 10.3389/fdata.2023.1229252

OPEN ACCESS

EDITED AND REVIEWED BY
Murat Kantarcioglu,
The University of Texas at Dallas, United States

*CORRESPONDENCE
Kirsi Marjaana Helkala
✉ kirsi.helkala@gmail.com

RECEIVED 26 May 2023
ACCEPTED 06 June 2023
PUBLISHED 19 June 2023

CITATION
Helkala KM, Lucas G, Barrett E and Syse H (2023) Editorial: Ethical challenges in AI-enhanced military operations.
Front. Big Data 6:1229252.
doi: 10.3389/fdata.2023.1229252

Editorial: Ethical challenges in AI-enhanced military operations

Kirsi Marjaana Helkala[1,2]*, George Lucas[3], Edward Barrett[4] and Henrik Syse[2]

[1]Norwegian Defence Cyber Academy, Norwegian Defence University College, Oslo, Norway, [2]Peace Research Institute Oslo, Oslo, Norway, [3]Department of Leadership, Ethics and Law, U.S. Naval Academy, Annapolis, MD, United States, [4]Stockdale Center, U.S. Naval Academy, Annapolis, MD, United States

KEYWORDS

artificial intelligence (AI), military ethics, virtue ethics, conventional combat, cyber conflict, strategic planning

Editorial on the Research Topic
Ethical challenges in AI-enhanced military operations

This Research Topic primarily focuses on the people—military personnel throughout the command structure—who serve in combat settings with AI-enabled machines. In a battlespace where machine autonomy is increasingly assuming functions once restricted to human beings, maintaining clear lines of human responsibility is of paramount importance. Clarifying this issue should improve ethical instruction within military training and educational institutions, as well as change how AI developers design their technologies. In turn, this will render ethical guidelines better tailored to the battlefield scenarios military personnel will confront in the future.

This collection aims to yield moral guidelines for the variety of military uses of AI technology, primarily in three areas:

Conventional armed conflict/battlefield combat;
Cyber military operations and cyber conflict;
Strategic planning for war and data-driven battlefield management.

Additionally, these essays examine the impact on the competency and character of human operators, *inter alia*, through the lens of virtue ethics. That is, they focus on individual character and the cultivation of moral predispositions that empower us to act responsibly amid the challenges to personal and professional life increasingly posed by the use of artificial intelligence in cyber security, kinetic warfare, and intelligence and strategic planning.

Within cyber security, AI-based tools can be used in both offensive and defensive cyber applications, from malware detection, network intrusion, and phishing and spam detection to intelligent threats and tools for attacking AI models. AI-based systems and their use in the kinetic battlefield encompass autonomous vehicles, drones, and swarms. Finally, intelligence systems that enable the examination and analysis of large data sets and integration of inputs from a vast array of sources enable ever-more effective planning, battlefield management, surveillance, and development of data-driven strategies focused on defense and national security (as is currently happening in Ukraine).

This Research Topic of "Frontiers in Big Data" originated as part of a project ("Warring with Machines: Artificial Intelligence and the Relevance of Virtue Ethics") at the Peace Research Institute Oslo (PRIO) focused on the uses and ethical impact of AI in military settings and special operations. Most of the papers were refereed and presented initially at international conferences

held in Rome (and co-organized by PRIO and Notre Dame University's Technology Ethics Center), at the McCain Conference in Annapolis, USA (organized by the Stockdale Center at the US Naval Academy in association with PRIO), as well as at annual conferences of the International Society of Military Ethics in Europe and the USA. Others were received in response to an international CFP through the Loop Science Network.

The papers encompass four research areas, each addressing the challenges of AI-enhanced military operations: (1) Artificial Intelligence and the Ethics of Warfare; (2) The Impact of Military Reliance on AI upon Human War Fighters and Their Control of Warfare; (3) Dignity and Respect; and (4) Gender Bias in Narratives of War.

The first topic includes a paper published separately from this Research Topic by Frontiers, namely, Regan and Davidovic (2023) "Just Preparation for War and AI-Enabled Weapons", as well as the papers "*The Comparative Ethics of Artificial-intelligence Methods for Military Applications*" by Rowe, and "*The PRC Considers Military AI Ethics: Can Autonomy be Trusted?*" by Metcalf. All three address AI in warfare from a military ethics perspective, addressing proper preparation and testing, the ethics of different algorithms in use, and the deeply political way in which the military ethics of AI is understood in China.

The second topic is addressed by Hovd in the paper "*Tools of War and Virtue-Institutional Structures as a Source of Ethical Deskilling*", which analyses military virtues as a species of moral virtues mediated by institutional and technological structures, meaning that professional roles and institutional structures are constitutive parts of what makes these virtues what they are, and thus the most likely source of ethical deskilling. The paper "*On the Purpose of Meaningful Human Control*" by Davidovic critically discusses and analyses calls for proper control of AI-enabled weapons systems, focusing on the *purpose* of such control, while "*The Ethics of AI-assisted War Fighter Enhancement Research and Experimentation*" by Moreno et al. discusses the problem of AI-wired war fighters facing affronts to cognitive liberty and to psychological and physiological health, as well as obstacles to integrating into military and civil society during their service and upon discharge, emphasizing the importance of ethics in the research underlying the use of such technologies.

The third topic includes the paper "*Resolving Responsibility Gaps for Lethal Autonomous Weapon Systems*" by Taylor Smith, in which the author suggests an understanding of collective responsibility for AI outcomes that can help resolve the "problem of many hands" and "responsibility gaps". It also includes Kahn's "*Lethal Autonomous Weapons Systems and Respect for Human Dignity*", which discusses whether actions involving the use of AI-enhanced weapons can respect human dignity. Kahn suggests criteria for the possibility of answering that question in the affirmative. Finally, "*Lethal Autonomous Weapons Systems, Revulsion, and Respect*" by Dean raises the issue of respect for public opinion and conventional attitudes as one develops lethal autonomous weapons systems, those opinions and attitudes often being very skeptical toward the development of such weaponry.

The fourth and last topic is addressed by Fisher in "*The Role of Gender in Providing Expert Advice on Cyber Conflict and Artificial Intelligence to Military Personnel*". Fisher argues that, as the role of cyber and AI grows in military operations, there is a need for military institutions to take gender into account in both training and policy, not least due to the fact that gender stereotypes attach to the role of cyber-engineer.

Together, these papers focus on the impact of current and anticipated military uses of AI-augmented technologies within the framework of military ethics and the just war tradition, as well as unique individual moral challenges that such technologies present. The effect of AI-augmentation on conceptions of human dignity, respect, and gender roles in military settings is found to be particularly problematic. Specific responses and remedies to the major problems described are proposed where feasible.

Author contributions

Each contributing author to this collection participated in conceptualization of their topic, writing of the original draft and subsequent revisions, and funding acquisition for the overall project. The editors approved the final submitted version.

Acknowledgments

We express our sincere gratitude to all the authors who proposed their work, all the researchers who took care to provide their most constructive comments and suggestions, and to the Frontiers team for their support. We are grateful for the support of the Research Council of Norway that contributed funding (for the "Warring with Machines" project) toward the publication of this collection as an e-Book.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Regan, M., Davidovic, J. (2023). Just preparation for war and AI-enabled weapons. *Front. Big Data*. 6, 1020107. doi: 10.3389/fdata.2023.1020107.

frontiers | Frontiers in Big Data

TYPE Hypothesis and Theory
PUBLISHED 09 September 2022
DOI 10.3389/fdata.2022.999293

OPEN ACCESS

EDITED BY
Edward Barrett,
United States Naval Academy,
United States

REVIEWED BY
Alec Walen,
Rutgers, The State University of New Jersey, United States
Jai Galliott,
University of Oxford, United Kingdom

*CORRESPONDENCE
Leonard Kahn
lakahn@loyno.edu

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 20 July 2022
ACCEPTED 17 August 2022
PUBLISHED 09 September 2022

CITATION
Kahn L (2022) Lethal autonomous weapon systems and respect for human dignity.
Front. Big Data 5:999293.
doi: 10.3389/fdata.2022.999293

Lethal autonomous weapon systems and respect for human dignity

Leonard Kahn*

College of Arts and Sciences, Loyola University New Orleans, New Orleans, LA, United States

Much of the literature concerning the ethics of lethal autonomous weapons systems (LAWS) has focused on the idea of human dignity. The lion's share of that literature has been devoted to arguing that the use of LAWS is inconsistent with human dignity, so their use should be prohibited. Call this position "Prohibitionism." Prohibitionists face several major obstacles. First, the concept of human dignity is itself a source of contention and difficult to operationalize. Second, Prohibitionists have struggled to form a consensus about a property P such that (i) all and only instances of LAWS have P and (ii) P is always inconsistent with human dignity. Third, an absolute ban on the use of LAWS seems implausible when they can be used on a limited basis for a good cause. Nevertheless, my main purpose here is to outline an alternative to Prohibitionism and recognize some of its advantages. This alternative, which I will call "Restrictionism," recognizes the basic intuition at the heart of Prohibitionism - namely, that the use of LAWS raises a concern about human dignity. Moreover, it understands this concern to be rooted in the idea that LAWS can make determinations without human involvement about whom to target for lethal action. However, Restrictionism differs from Prohibitionism in several ways. First, it stipulates a basic standard for respecting human dignity. This basic standard is met by an action in a just war if and only if the action conforms with the following requirements: (i) the action is militarily necessary, (ii) the action involves a distinction between combatants and non-combatants, (iii) noncombatants are not targeted for harm, and (iv) any and all incidental harm to non-combatants is minimized. In short, the use of LAWS meets the standard of basic respect for human dignity if and only if it acts in a way that is functionally isomorphic with how a responsible combatant would act. This approach leaves open the question of whether and under what conditions LAWS can meet the standard of basic respect for human dignity.

KEYWORDS

military ethics, AI ethics, lethal autonomous weapon systems, artificial intelligence, applied ethics

Much of the literature concerning the ethics of lethal autonomous weapons systems (or LAWS as they are usually called) has focused on the idea of human dignity, and the lion's share of that literature has been devoted to arguing that (1) the use of LAWS without meaningful human control violates human dignity, so (2) their use is morally prohibited and, therefore, (3) should be

illegal.[1] Peter Asaro provides an admirably clear statement of this view when he writes,

> As a matter of the preservation of human morality, dignity, justice, and law we cannot accept an automated system making the decision to take a human life. And we should respect this by prohibiting autonomous weapon systems (Asaro, 2012).[2]

Call this view "Prohibitionism."

There is much to be said in favor of Prohibitionism. It proceeds from intuitive premises, and it has clear policy implications that align well with general concerns about limiting the use of lethal violence in war. However, I take a different but related approach in this paper. In particular, I sketch an underexplored alternative to Prohibitionism and briefly offer considerations in favor of it. I conclude by suggesting some implications for the development of LAWS.

In order to understand this alternative view a little better, it will be helpful to take a step back and get a better look at the broader conceptual landscape. To that end, consider two claims that will be of central importance throughout my discussion:

- The Permissibility Claim: With respect to considerations of human dignity, it is morally *permissible* to use LAWS.
- The Requirement Claim: With respect to considerations of human dignity, it is morally *required* to use LAWS.[3]

There are a number of ways to mix and match evaluations of the Permissibility Claim and the Requirement Claim, but only two are important here. First, Prohibitionists are committed to the view that both the Permissibility Claim and the Requirement Claim are always false. Second, the alternative that I'll explore maintains that both the Permissibility Claim and the Requirement Claim are sometimes true and sometimes false, depending on context, particularly the level of technological development. I'll call this alternative "Weak Restrictionism."

More about Weak Restrictionism in a moment. I owe the reader a few words about the idea of human dignity, though I can treat this topic only superficially in the space available to me. It is important to begin by acknowledging frequent complaints that human dignity is a "useless concept" (Macklin, 2003) or a "squishy, subjective notion, hardly up to the heavyweight moral demands assigned to it" (Pinker, 2008). We should recall that complaints of this kind are not new or unique to discussion of LAWS. Michael Rosen reminds us that Schopenhauer denounced talk of the "dignity of man" as "the shiboleth of all the perplexed and empty-headed moralists who conceal behind that imposing expression their lack of any real basis of morals, or, at any rate, of one that had meaning" (Schopenhauer, 1998; quoted in Rosen, 2012, p. 1).[4]

Of course, it's easy to caricature and then ridicule the idea of human dignity. This is a tempting mistake - but a mistake nonetheless. While there's no doubt both that the concept is not fully understood (Kateb, 2014) and that competing interpretations of it are often difficult to operationalize (Polonko and Lombardo, 2005), we can try to make some progress.

For the purposes of this paper, I understand human dignity as a status concept.[5] A little more precisely, human dignity concerns the moral standing that human beings have in virtue of the intrinsically valuable features that characterize most mature members of our species. According to this way of thinking, human dignity can play the role of justifying international humanitarian law (Regan, 2019, p. 213) and international human rights law (Luban, 2015, p. 270). Actions and attitudes that treat human beings as if they lacked these valuable qualities are threats to human dignity since they treat us in ways that are humiliating.[6] For now, let me note that human dignity so understood need not be based on "absolute, unconditional, and incomparable value or worth" (Parfit, 2011, p. 239). The view that human dignity has no rivals with regard to the value on which it is based—a view associated with Kant (Dean, 2006, p. 37)—is consistent with what I will say here, though it is not entailed by it. All that matters for the purposes of this paper is that the phrase "human dignity" denotes a *high* moral status that imposes

1 I will be concerned in this paper only with the use of LAWS *without* meaningful human control since their use with meaningful human control - i.e., without full autonomy - is less interesting and involves complications and qualifications that I cannot address here. In order to avoid tedium, I will drop the phrase "without meaningful human control" in the rest of this paper, but I take it as understood. On meaningful human control, see Santoni de Sio and Van den Hoven (2018).

2 See also, e.g., Docherty (2014), Sparrow (2016), Heyns (2017), Rosert and Sauer (2019), Sharkey (2019), Sauer (2021), and Saxon (2021). I look at some arguments offered in favor of Prohibitionism later in this paper. For earlier criticism, see Birnbacher (2016).

3 Obviously, the domain of agents affected by the Requirement Claim is limited to those who can use LAWS since "ought," in some suitable sense, entails "can" (Griffin, 2010). I understand permitted and required in the standard way these terms are mutually defined in deontic logic. X is permitted to do F if and only if X is not required not to F, and X is required to do F if and only if X is not permitted not to do X (Føllesdal and Hilpinen, 1970; Rönnedal, 2010, p. 29).

4 Rosen's characterization of Schopenhauer as the "the Ebeneezer Scrooge of nineteenth-century philosophy," is too good to leave unmentioned.

5 Luban (2007, p. 89) and Lysaker and Syse (2016, p. 117–118).

6 On human dignity as non-humiliation, see Margalit (1996), Jaber (2000), Nussbaum (2009), Luban (2009), Killmister (2010), Sharkey (2014), Vorhaus (2015), Coghlan (2018), and Gisbertz (2018).

significant moral restrictions on certain classes of actions and attitudes. This is not the place to enumerate these features or explain why they are valuable. Such features, while perhaps not unknowable, are certainly not easily known and require a separate line of inquiry (Barrett, 2013, p. 5). Rather, I will fall in line with a tradition of grounding to at least some degree human dignity in personal autonomy, our capacity to govern ourselves in light of our values, suitably understood (e.g., Oshana, 2016).

Now that we have a barebones account of human dignity let's work our way back to Prohibitionism and Weak Restrictionism by means of the Principle of Discrimination, which requires us both to distinguish between enemy combatants and non-combatants and to avoid targeting the latter for harm.[7] Some Prohibitionists connect the idea that the use of LAWS is a violation of human dignity with the idea that their use transgresses against the Principle of Discrimination (Gubrud, 2014).

Their use transgresses against the Principle of Discrimination, on this line of thinking, for at least two reasons. First, the sensory systems of LAWS are incapable of distinguishing reliably between combatants and non-combatants, because there is no algorithm for determining whether someone is a combatant. Or, as Sharkey puts it, LAWS "do not have adequate sensory or vision processing systems for separating combatants from civilians, particularly in insurgent warfare, or for recognizing wounded or surrendering combatants" (2012, p. 288). Second, LAWS lack higher-order situational awareness that would promote the capacity to make this distinction (Sharkey, 2019, p. 76). These are, I acknowledge, just two complaints. One might add that LAWS are likely to have a hard time recognizing attempts to surrender. But this will do.

This, if true, shows that the use of LAWS could be a violation of the Principle of Discrimination. But a further argument would be necessary to show that this violation would also constitute an offense against human dignity. Satisfactory arguments of this sort can be harder to find, but a promising start can be found with the observation that actions that ignore the distinction between combatants and non-combatants thereby fail to recognize and treat appropriately the value of the personal autonomy of those who have chosen not to become combatants and who have, therefore, rendered themselves essentially defenseless against the use of armed force.[8] If we transgress against the Principle of Discrimination, then we treat humans as if they were non-humans who cannot make the choice to refrain from harming others by not engaging in military service and, in the process, rendering themselves vulnerable—indeed mortally vulnerable—to others.[9] In doing so, we blatantly ignore one of the characteristic features of our species—personal autonomy—that makes our lives intrinsically valuable.[10]

I think we should grant the Prohibitionist the claim that violations of the Principle of Discrimination are infringements of human dignity in more or less the way that I've just suggested. And I think we should also acknowledge that the use of LAWS can be a violation of the Principle of Discrimination, especially given their fairly crude level of development at the moment. Hence, it is at least sometimes the case that the Permissibility Claim is false. But even if we concede that the use of the fairly primitive LAWS now available would be a violation of the Principle of Discrimination, it does not follow that the same will be true of the use of future LAWS. Along those lines, consider a somewhat complicated scenario I will call

> Case 1: Alfastan is fighting a just war against Betaville. Gamma, an officer in Alfastan's military, has the opportunity to capture Point Delta, which is necessary to Alfastan's effort to accomplish its just war aims. However, the only soldiers available to Gamma at the moment are in Company Epsilon. Gamma believes that these warfighters will violate the Principle of Discrimination if they are ordered to take Point Delta. Nevertheless, Gamma can instead use highly sophisticated LAWS to take Point Delta. These futuristic LAWS can distinguish between combatants and non-combatants with a high degree of accuracy, thereby vastly lowering the risk of a violation of the Principle of Discrimination.[11]

It seems to me that it is permissible on the basis of considerations of human dignity for Gamma to use their LAWS

7 The devil is in the details when it comes to the Principle of Discrimination, as is so often the case. See Nagel (1972), Walzer (1977, p. 160–175), McMahan (2009, p. 11–12), and Frowe (2015, p. 82–83). However, discussion of these details can be postponed until another time, and the flatfooted distinction I make here between combatants and noncombatants can be nuanced as necessary.

8 The fact that non-combatants have chosen not to make themselves liable to attack does not entail that all combatants have chosen to do so. Whether all combatants are indeed liable to attack is not something that needs to be resolved in this paper. I am grateful to an anonymous referee for pushing me to clarify this point.

9 Perhaps those who have been coerced into the role of combatant have a different status than those who have assumed the role in the absence of coercion. But the brevity of this paper requires painting with a broad brush.

10 For an alternative approach to human dignity and the Principle of Discrimination, see Kasher (2014).

11 An anonymous referee pointed out, reasonably enough, a similarity between Case 1 and some thought experiments in Strawser (2010). However, Strawser's focus is on neither autonomous weapons nor human dignity.

instead of the soldiers in Company Epsilon to take Point Delta. This fact weighs against understanding the Permissibility Claim as always false, as Prohibitionists maintain. Or, to put the matter a little differently, if my judgment is correct then it is sometimes permissible on the grounds of human dignity to use LAWS. Now consider

> Case 2: This case is similar to Case 1 except that by taking Point Delta, Gamma can prevent large-scale assaults on the basic human dignity of a group of non-combatants, which Gamma is certain will otherwise occur as a result of the actions by their adversary.

Plausibly, on the basis of considerations of human dignity, it is not only permissible but required for Gamma to use their highly sophisticated LAWS instead of their soldiers to take Point Delta in Case 2. That is to say, one would not be permitted not to use LAWS in this situation. Why? Only by using these LAWS, which do not themselves violate the Principle of Discriminaiton, can Gamma prevent the large-scale assaults on the human dignity of many non-combatants. If Gamma is required to do so, then there are at least some situations in which the Requirement Claim is true, a fact which is consistent with Weak Restrictionism but not with Prohibitionism.

It might be objected that the two thought experiments I've offered rely too heavily on warfighters being incapable or unwilling to act on the Principle of Discriminaiton. So consider what I will call the maximalist extension of the Principle of Discrimination: If one can employ several means M_1, M_2, … M_n, etc. to achieve an otherwise morally legitimate military objective, and one of the means, M_i is more likely than any of the other means to distinguish non-combatants from enemy combatants while not targeting the former for harm, then there is a pro tanto reason to use M_i rather than any other means.[12] The reason to favor M_i is only prima facie since there might be countervailing reasons that favor another means. Nevertheless, if the maximalist extension of the Principle of Discrimination is correct,[13] then there are more possible circumstances in which the use of LAWS is not only permissible but required by considerations of human dignity. Consider

> Case 3: The case is similar to Case 2. However, Gamma can also use Company Zeta to take Point Delta (to avoid the widespread assaults on human dignity), where Company Zeta, unlike Company Epsilon, would be unlikely to violate the Principle of Discrimination in the process. However, Gamma could instead use a highly advanced form of LAWS to take Point Delta. This LAWS is even more likely than the members of Company Zeta to distinguish accurately between enemy combatants and non-combatants and to avoid targeting non-combatants for harm.

In the absence of any countervailing considerations, I think Gamma is required by considerations of human dignity to use LAWS in Case 3 since it will achieve the same end of avoiding assaults on human dignity and do so by employing a means that is less likely to be a violation of human dignity.

More generally, Case 3 points to a surprising conclusion. Even in circumstances where it is possible meet *human* standards of discrimination, the continued development of technology and artificial intelligence might make it the case that we are required nevertheless to use LAWS precisely because doing so is more reliably discriminate. Warfighters might "be all they can be" yet still not be enough to do what they are required to do. In such cases it might be that human moral virtue must take a back seat to inhuman technological excellence.

Let me turn now to another suggestion made by Prohibitionists about how the use of LAWS is a threat to human dignity. This suggestion is more difficult to articulate adequately than the previous one, and it has many critics. My aim here is not to demonstrate its truth; it is to show that, even if we conceded to Prohibitionists that LAWS are a threat to human dignity in this way, Weak Restrictionism is still a more plausible position than Restrictionism. Greg Reichberg and Henrik Syse get us off to a good start with regard to this suggestion when they propose the possibility that "To be killed by machine decision would debase warfare into mere slaughter, as though the enemy combatant were on a par with an animal killed on an automated conveyor belt" (Reichberg and Syse, 2021, p. 153).[14] However, it might get slightly closer to the view of many Prohibitionists to say that the use of LAWS debases not warfare but rather the human dignity of those who are caught up in it.

In order to get a better grip on this slippery idea, imagine (and here I betray my roots as a full-time college administrator) that your university has just purchased a derelict building that it plans on tearing down and replacing with a new dormitory for the university's students. However, the derelict building is infested with vermin, who are carriers of a lethal disease. While it is possible to send in workers to eradicate the pests, it is also possible to send in an autonomous robot to do the job. All other things being equal, many of us would think that sending in the autonomous robot is no worse—and possibly much better—than sending in workers to do the same. The thought here is that

12 To be sure if there are two means M_j and M_k that are equally likely than another of the other means to distinguish between enemy combatants and noncombatants while not to targeting the latter for harm, and no other means are more likely than these, than there is a pro tanto reason to use either M_j or M_k.

13 One way to vindicate the maximalist extension of the Principle of Discrimination is by means of the Doctrine of Double Intention. See Lee (2004) and Zohar (2007).

14 See also, e.g., Goose and Wareham (2016), Leveringhaus (2018), and Rosert and Sauer (2019).

when it comes to eliminating vermin, there is nothing wrong *per se* about allowing lethal determinations to be made by non-humans. But many of us balk when it comes to permitting these lethal determinations to be made by non-humans when it is *human* lives on the line. The thought continues that, even if the autonomous robot is just as accurate as a human being concerning who is to be targeted with lethal force, the use of the autonomous robot transforms an act of war into something akin to the slaughter of an unwanted rodent. The use of an autonomous robot, such as a LAWS, expresses, the thought concludes, an attitude of contempt toward humans that amounts to an assault on human dignity. Note that the point here is not that the deaths of those killed by LAWS would be more painful, more protracted, or more gruesome. It might or might not be any of these things, but these issues are orthogonal to the point at hand - namely, that the action expresses contempt for the combatants and their value as human beings.

Before commenting on this suggestion, I will need to make two further points about the nature of human dignity. First, it is possible to distinguish two distinct levels of human dignity. This distinction is easier to see when considering concrete examples of human dignity being violated. Begin with the plausible thought that being denied certain goods that are not central to the exercise of personal autonomy over the course of an entire life can count as a violation of one's human dignity. For example, suppose that your sexual orientation is not those of the majority of the people among whom you live.

Nevertheless, I think that there is a significant qualitative difference between violations of human dignity of this sort and violations associated with, for instance, torture (Luban, 2007, p. 162–204), rape (Nussbaum, 2009), and enslavement (Hörnle, 2012). In cases such as these, the contravention of human dignity is a thing apart from the cases of bigotry described above, loathsome though they are. I can offer no more than a promissory note here to explain this difference in greater detail, but I think it will do for our present purposes to distinguish between basic human dignity and non-basic human dignity. Violations of basic human dignity are those that have a significant negative effect on our most fundamental abilities as humans to live self-directed lives within the bounds of morality. To be sure, this distinction between basic and non-basic human dignity deserves more care and can be drawn with considerably more nuance than I have allowed myself here. Nevertheless, a rough-and-ready version of the distinction will be enough for the moment.

Here is the second point I want to make about human dignity. Addressing issues of basic human dignity is considerably more urgent than addressing issues of non-basic human dignity. Rather dogmatically, I will say that the two forms of human dignity have something approximating a lexical relationship with regard to their moral importance. Let me spell out in some detail what I mean by that claim. Consider an agent A who has in their power the ability by acting to affect a group of stakeholders S_1 through S_n. Imagine that if A can do either one of two actions, a_1 or a_2. And further imagine that if A does a_1, then they will ensure that at least one of S_1 through S_n does not suffer a violation of their basic human dignity. However, if A does a_2, then A will not ensure that at least one of S_1 through S_n does not suffer a violation of their basic human dignity, though they will bring it about that at least one of S_1 through S_n does not suffer a violation of their non-basic human dignity. Let the Priority Hypothesis be that in these conditions A is required to do a_1, even though doing so will allow for a violation of non-basic human dignity among the stakeholders.

Even in its current form, the Priority Hypothesis provides support for Weak Restrictionism. In order to see why this is the case, return to Case 1. Assume for the sake of argument that the use of Epsilon Company to take Point Delta does not express contempt for the human dignity of the combatants who would defend it but, consistent with the suggestion of the Prohibitionists, that the use of LAWS would do so. But recall from this thought experiment that the LAWS will also more reliably distinguish between enemy combatants and non-combatants and while avoiding targeting non-combatants for harm than the other options available to Gamma. I think it is plausible to say that disregarding the opportunity to safeguard non-combatants is most likely a violation of basic human dignity, while expressing contempt toward combatants is more credibly a violation of non-basic human dignity. An act that fails to be as discriminate as possible certainly has a significant negative effect on our most fundamental abilities as humans to live self-directed lives within the bounds of morality since it will, for that very reason, result in injuries and death to those who have chosen not to be combatants. However, an act that expresses contempt toward combatants will not, *per se*, have a negative impact of this kind. The combatants who are injured and killed by the LAWS will not be diminished as humans by the attitude expressed by the use of the LAWS. The expression itself will not prevent them developing and promoting the valuable properties that are characteristic of being human. But the Priority Hypothesis requires us to give greater priority to avoiding the violation of basic human dignity than the violation of non-basic human dignity.

To say all of this, of course, is not to endorse or to trivialize such expressions of contempt. If Prohibitionists are correct that the use of LAWS does express disdain for human dignity, then there will be many situations in which it is impermissible to use them. However, Weak Restrictionists accept this point; indeed, they insist on it. They differ from Prohibitionists in virtue of also accepting the claims that it is sometimes permissible and even required to use LAWS because of considerations of human dignity. And it appears that in Case 1, the use of LAWS is required even if we grant the point that their use involves a violation of non-basic human dignity. Indeed, that is just what Weak Restrictionists have in mind when they assert that both the Permissibility Claim and the Requirement Clam are sometimes

true. One need not throw up one's hands about the possibility of dignified death in war to acknowledge this (Scharre, 2018, p. 288). Nor is it necessary to hold out the hope that "one can maintain dignity in the face of indignity" (Young, 2021, p. 173). Weak Restrictionists need only maintain considerations of human dignity sometimes favor the use of LAWS all things considered, even if they do not do so unambiguously because it is sometimes impossible to act in ways that have no negative consequences for human dignity. The use of LAWS would not be unique among military actions in being ethically ambiguous for this reason. Something like this might also be true when it comes to, for example, the ethics of military intelligence (Bailey and Galich, 2012, p. 86).

One of the conceits of this paper is that it is possible that there will be LAWS that are better than humans—perhaps far better than humans—with respect to distinguishing combatants from non-combatants in the battle space. Implicitly, I've assumed that LAWS can do many other things better than humans can to avoid violations of human dignity as well. These ideas are far from new (Arkin, 2010). However, they are also far from being true of our world as I write this sentence in July 2022. Currently, there are few if any contexts in which it would be reasonable to expect LAWS consistently to perform better than well-trained and conscientious human combatants with respect to discrimination. Moreover, it is impossible to say with certainty when (or perhaps even *if*) LAWS will become robust and competent enough to be deployed in combat without meaningful human control. So it's natural to wonder whether introducing the possibility of futuristic LAWS does anything to advance the conversation.

Let me conclude by explaining why all of this—I think—matters. If Prohibitionists are correct, then considerations of human dignity require us to halt research and development into LAWS. Since they cannot be used in a manner that is consistent with human dignity, they should not be developed, and all of our energy should go into banning their use. Recall from the beginning of this paper Peter Asaro's insistence that "we should respect this [i.e., human dignity] by prohibiting autonomous weapon systems." But if Weak Restrictionists are correct, then considerations of human dignity actually require further research into and development of LAWS since there are possible circumstances in which human dignity demands their use. Furthermore, if Weak Restrictionists are right, then considerations of human dignity at least partially ought to set the agenda for how LAWS are to be developed. It also ought to inform any and all attempts to regulate the development and use of LAWS at the international level. So the implications of the two views are very different, and the fact that the LAWS I've used in thought experiments throughout this talk are not yet extant is no objection.[15]

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

[15]I am very grateful to attendees of the 2022 McCain Conference at the US Naval Academy and the 2022 meeting of the International Society of Military Ethics at the University of Colorado, Colorado Spring as well as to two anonymous referees from their comments, questions, and criticisms. Readers of this paper who have spent the 2021–2022 academic year with me in the Stockdale Seminar will recognize the degree to which I am indebted to its members for shaping - and reshaping - my thinking on this topic. I am deeply grateful to them for this. I have attempted to register these debts in the parenthetical references and footnotes, though I doubt I have managed to do so in every case where I should have. Whatever flaws remain in the paper are my own.

References

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *J. Milit. Ethics* 9, 332–341. doi: 10.1080/15027570.2010.536402

Asaro, P. (2012). On banning autonomous lethal systems: human rights, automation and the dehumanizing of lethal decision-making, *Int. Rev. Red Cross* 94, 687–709. doi: 10.1017/S1816383112000768

Bailey, C. E., and Galich, S. (2012). Code of ethics: the intelligence community. *Int. J. Intell. Ethics* 3, 77–99. Available online at: https://journals.flvc.org/ijie/article/view/83454

Barrett, E. (2013). Warfare in a new domain: the ethics of military cyber-operations. *J. Milit. Ethics* 12, 4–17. doi: 10.1080/15027570.2013.782633

Birnbacher, D. (2016). "Are autonomous weapons systems a threat to human dignity?", in *Autonomous Weapons Systems: Law, Ethics, Policy*, eds N. Buhta, S. Beck, R. Geiβ, H.-Y. Liu, C. Kreβ (Cambridge: Cambridge University Press), 105–121.

Coghlan, S. (2018). The moral depth of human dignity. *Philos. Investig.* 41, 70–93. doi: 10.1111/phin.12177

Dean, R. (2006). *The Value of Humanity in Kant's Moral Theory*. Clarendon: Oxford University Press.

Docherty, B. L. (2014). *Shaking the Foundations: The Human Rights Implications of Killer Robots*. Available online at: https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots (accessed May 3, 2022).

Føllesdal, D., and Hilpinen, R. (1970). "Deontic logic: an introduction," in *Deontic Logic: Introductory and Systematic Readings*, ed R. Hilpinen (Dordrecht: Springer), 1–35.

Frowe, H. (2015). *The Ethics ofWar and Peace*. 1st Edition. Oxford: Routledge.

Gisbertz, P. (2018). Overcoming doctrinal school thought: a unifying approach to human dignity. *Ratio Juris* 31, 196–207. doi: 10.1111/raju.12204

Goose, S., and Wareham, M. (2016). The growing international movement against killer robots. *Harvard Int. Rev.* 37, 28–34. Available online at: https://www.hrw.org/news/2017/01/05/growing-international-movement-against-killer-robots

Griffin, J. (2010). *Ought Implies' Can'*. Lindley Lecture. University of Kansas, Department of Philosophy. Available online at: https://kuscholarworks.ku.edu/bitstream/handle/1808/12422/Ought%20Implies%20Can-2010.pdf?sequence$=$1andisAllowed$=$y (accessed May 3, 2022).

Gubrud, M. (2014). Stopping killer robots. *Bull. Atom. Sci.* 70, 32–42. doi: 10.1177/0096340213516745

Heyns, C. (2017). Autonomous weapons in armed conflict and the right to a dignified life: an African perspective. *S. Afr. J. Hum. Rights* 33, 46–71. doi: 10.1080/02587203.2017.1303903

Hörnle, T. (2012). Criminalizing behavior to protect human dignity. *Crim. Law Philos.* 6, 307–325. doi: 10.1007/s11572-012-9177-6

Jaber, D. (2000). Human dignity and the dignity of creatures. *J. Agric. Environ. Ethics* 13, 29–42. doi: 10.1007/BF02694133

Kasher, A. (2014). Combatants' life and human dignity. *Isr. Yearb. Hum. Rights* 44, 219–246. doi: 10.1163/9789004423954_007

Kateb, G. (2014). *Human Dignity*. Cambridge, MA: Harvard University Press.

Killmister, S. (2010). Dignity: not such a useless concept. *J. Med. Ethics* 36, 160–164. doi: 10.1136/jme.2009.031393

Lee, S. (2004). Double effect, double intention, and asymmetric warfare. *J. Milit. Ethics* 3, 233–251. doi: 10.1080/15027570410006183

Leveringhaus, A. (2018). What's so bad about killer robots? *J. Appl. Philos.* 35, 341–358. doi: 10.1111/japp.12200

Luban, D. (2007). *Legal Ethics and Human Dignity*. Cambridge.

Luban, D. (2009). Human dignity, humiliation, and torture. *Kennedy Inst. Ethics J.* 19, 211–230. doi: 10.1353/ken.0.0292

Luban, D. (2015). "Human rights pragmatism and human dignity," in *Philosophical Foundations of Human Rights*, eds C. Rowan, S. Matthew Liao, and M. Renzo (Oxford: Oxford University Press), 263–278.

Lysaker, O., and Syse, H. (2016). The dignity in free speech: Civility norms in post-terror societies. *Nord. J. Hum. Rights* 34, 104–123. doi: 10.1080/18918131.2016.1212691

Macklin, R. (2003). Dignity is a useless concept. *Br. Med. J.* 327, 1419–1420. doi: 10.1136/bmj.327.7429.1419

Margalit, A. (1996). *The Decent Society*. Cambridge, MA: Harvard.

McMahan, J. (2009). *Killing in War*. Oxford.

Nagel, T. (1972). War and massacre. *Philos. Public Affairs* 1, 123–144.

Nussbaum, M. (2009). "Human dignity and political entitlements," in *Human Dignity and Bioethics*, eds E. Pellegrino, A. Schulman, and T. Merrill (Southbend, IN: University of Notre Dame Press), 245–263.

Oshana, M. (2016). *Personal Autonomy in Society*. Oxford: Routledge.

Parfit, D. (2011). *On What Matters*, *Volume* 1. Oxford.

Pinker, S. (2008). *The Stupidity of Dignity*. New Republic, 28–31. Available online at: https://newrepublic.com/article/64674/the-stupidity-dignity

Polonko, K. A., and Lombardo, L. X. (2005). Human dignity and children: operationalizing a human rights concept. *Glob. Bioethics* 18, 17–35. doi: 10.1080/11287462.2005.10800863

Regan, M. (2019). From protecting lives to protecting states: use of force across the threat continuum. *J. Natl. Sec. Law Policy* 10, 171–236. Available online at: https://heinonline.org/HOL/Page?handle=hein.journals/jnatselp10&div=10&g_sent=1&casa_token=8IAntyxooy0AAAAA:GuoPWy35NRowbo7RrDzXSkWIige80sUdf-NDYdkc9bKOC4q37-hAM0Nd4vsktErfhWwPnoIh&collection=journals

Reichberg, G., and Syse, H. (2021). *Applying AI on the Battlefield: The Ethical Debates*. Springer, 147–159. Available online at: https://link.springer.com/content/pdf/10.1007/978-3-030-54173-6.pdf

Rönnedal, D. (2010). *An Introduction to Deontic Logic*. Stockholm: Creative Space.

Rosen, M. (2012). *Dignity: Its History and Meaning*. Cambridge, MA: Harvard.

Rosert, E., and Sauer, F. (2019). Prohibiting autonomous weapons: put human dignity first. *Global Policy* 10, 370–375. doi: 10.1111/1758-5899.12691

Santoni de Sio, F., and Van den Hoven, J. (2018). Meaningful human control over autonomous systems: a philosophical account. *Front. Robot. AI* 28, 1–15. doi: 10.3389/frobt.2018.00015

Sauer, F. (2021). *Lethal Autonomous Weapons Systems*. Oxford: Routledge, 237–250.

Saxon, D. (2021). *Fighting Machines: Autonomous Weapons and Human Dignity*. Philadelphia, PA: University of Pennsylvania.

Scharre, P. (2018). *Army of None: Autonomous Weapons and the Future of War*. New York, NY: Norton.

Schopenhauer, A. (1998). *On the Basis of Morality*. Indianapolis, IN: Hackett.

Sharkey, A. (2014). Robots and human dignity: a consideration of the effects of robot care on the dignity of older people. *Ethics Inf. Technol.* 16, 63–75. doi: 10.1007/s10676-014-9338-5

Sharkey, A. (2019). Autonomous weapons systems, killer robots and human dignity. *Ethics Inf. Technol.* 21, 75–87. doi: 10.1007/s10676-018-9494-0

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

Strawser, B. J. (2010). Moral predators: the duty to employ uninhabited aerial vehicles. *J. Milit. Ethics* 9, 342–368. doi: 10.1080/15027570.2010.536403

Vorhaus, J. (2015). Dignity, capability, and profound disability. *Metaphilosophy* 46, 462–478. doi: 10.1111/meta.12141

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

Young, G. (2021). On the indignity of killer robots. *Ethics Inf. Technol.* 23, 473–482. doi: 10.1007/s10676-021-09590-2

Zohar, N. (2007). Double effect and double intention: a collectivist perspective. *Israel Law Rev.* 40, 730–742. doi: 10.1017/S0021223700013534

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 09 September 2022
DOI 10.3389/fdata.2022.978734

OPEN ACCESS

EDITED BY
George Lucas,
United States Naval Academy,
United States

REVIEWED BY
Patrick Lin,
California Polytechnic State University,
United States
Pauline Shanks Kaurin,
Naval War College, United States

*CORRESPONDENCE
Blake Hereth
Blake_Hereth@uml.edu

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 26 June 2022
ACCEPTED 08 August 2022
PUBLISHED 09 September 2022

CITATION
Moreno J, Gross ML, Becker J, Hereth B, Shortland ND III and Evans NG (2022) The ethics of AI-assisted warfighter enhancement research and experimentation: Historical perspectives and ethical challenges. *Front. Big Data* 5:978734. doi: 10.3389/fdata.2022.978734

The ethics of AI-assisted warfighter enhancement research and experimentation: Historical perspectives and ethical challenges

Jonathan Moreno[1], Michael L. Gross[2], Jack Becker[3], Blake Hereth[4]*, Neil D. Shortland III[5] and Nicholas G. Evans[4]

[1]Department of Bioethics, School of Medicine, University of Pennsylvania, Philadelphia, PA, United States, [2]School of Political Science, University of Haifa, Haifa, Israel, [3]Harvard Law School, Cambridge, MA, United States, [4]Department of Philosophy, University of Massachusetts at Lowell, Lowell, MA, United States, [5]School of Criminology and Justice Studies, University of Massachusetts at Lowell, Lowell, MA, United States

The military applications of AI raise myriad ethical challenges. Critical among them is how AI integrates with human decision making to enhance cognitive performance on the battlefield. AI applications range from augmented reality devices to assist learning and improve training to implantable Brain-Computer Interfaces (BCI) to create bionic "super soldiers." As these technologies mature, AI-wired warfighters face potential affronts to cognitive liberty, psychological and physiological health risks and obstacles to integrating into military and civil society during their service and upon discharge. Before coming online and operational, however, AI-assisted technologies and neural interfaces require extensive research and human experimentation. Each endeavor raises additional ethical concerns that have been historically ignored thereby leaving military and medical scientists without a cogent ethics protocol for sustainable research. In this way, this paper is a "prequel" to the current debate over enhancement which largely considers neuro-technologies once they are already out the door and operational. To lay the ethics foundation for AI-assisted warfighter enhancement research, we present an historical overview of its technological development followed by a presentation of salient ethics research issues. We begin with a historical survey of AI neuro-enhancement research highlighting the ethics lacunae of its development. We demonstrate the unique ethical problems posed by the convergence of several technologies in the military research setting. Then we address these deficiencies by emphasizing how AI-assisted warfighter enhancement research must pay particular attention to military necessity, and the medical and military cost-benefit tradeoffs of emerging technologies, all

attending to the unique status of warfighters as experimental subjects. Finally, our focus is the enhancement of friendly or compatriot warfighters and not, as others have focused, enhancements intended to pacify enemy warfighters.

KEYWORDS

artificial intelligence, warfighter enhancement, human research, experimentation, super soldiers

Introduction

Since the turn of the century, the ethics of research on human performance enhancement in the civilian setting has become an area of vigorous scholarship, not only with regard to compliance with traditional ethical standards but also in light of developments in related fields like cognitive neuroscience that seeks to understand the structure of the human brain and cognition; and artificial intelligence (AI) that seeks to develop machines capable of performing tasks that would ordinarily require human cognition. These fields have promise to enhance human capacities and improve performance in a range of tasks, such as through the use of brain-computer interfaces (BCI) that connect humans to computers, potentially in both directions, and even brain-to-brain interfaces. These fields, moreover, are interrelated: Neuroscience benefits greatly from artificial intelligence to probe the human brain and create novel technologies to investigate and treat disease or enhance performance. For instance, applications include emotion suppression, enhanced awareness, WiFi capability, and the like. AI, meanwhile, benefits from an understanding of human cognition and neurology to develop better and "smarter" machines capable of acting autonomously. These convergent fields are particularly attractive to, for example, the defense industry, for the ability to combine the lateral thinking and instinct of warfighters with the processing power of AI.

The military applications of AI raise myriad ethical challenges across countries [e.g., (Australian DoD (Department of Defence), 2020; UK Ministry of Defence, 2021)]. Critical among them is how AI integrates with human decision making to enhance cognitive performance on the battlefield. AI applications range from augmented reality devices to assist learning and improve training to implantable BCI to create bionic "super soldiers." As these technologies mature, AI-wired warfighters face potential affronts to cognitive liberty, psychological and physiological health risks and obstacles to integrating into military and civil society during their service and upon discharge (Denning et al., 2009). Before coming online and operational, however, AI-assisted technologies and neural interfaces require extensive research and human experimentation. Each endeavor raises additional ethical concerns that have been historically ignored thereby leaving military and medical scientists without a cogent ethics protocol for sustainable research. In this way, this paper is a "prequel" to the current debate over enhancement which largely considers neuro-technologies once they are already out the door and operational (ICRC, 2006).

To lay the ethics foundation for AI-assisted warfighter enhancement research, we present an historical overview of its technological development followed by a presentation of salient ethics research issues. We begin with a historical survey of AI neuro-enhancement research highlighting the ethics lacunae of its development. We demonstrate the unique ethical problems posed by the convergence of several technologies in the military research setting. Then we address these deficiencies by emphasizing how AI-assisted warfighter enhancement research must pay particular attention to military necessity, and the medical and military cost-benefit tradeoffs of emerging technologies, all attending to the unique status of warfighters as experimental subjects. Finally, our focus is the enhancement of friendly or compatriot warfighters and not, as others have emphasized (Hereth, 2022), enhancements intended to pacify enemy warfighters.

Historical background of military artificial intelligence and neurotechnology

Artificial intelligence and defense planning

In 1956, computer scientist John McCarthy organized the Dartmouth Summer Research Conference where the term "artificial intelligence" was coined. McCarthy was frustrated that little had been written about the idea that computers could possess intelligence. The 1956 Dartmouth conference is regarded as the origin of the approach known affectionately, and sometimes critically, as "good old-fashioned AI" or GOFAI, which is built on symbolic reasoning and logic. The more recent framework that utilizes mathematical models or "neural networks" capable of searching for patterns in vast quantities of data is often called "connectionism" and produces machine-learning using algorithms. Despite its rich history and ubiquity in the modern world, there remain certain basic disagreements both about what "AI" really means and whether advances in

computing will ever lead to human-level intelligence or even a "superintelligence" that threatens human civilization. These disagreements about the definition and ultimate power of AI do not pose problems for this paper as our focus is on AI-*enabled* technologies, i.e., those that exploit systems that are generally regarded as based on principles of AI.

What can be said is that the Dartmouth conference established a fundamental assumption about the nature of intelligence itself, as a set of cognitive capacities directed toward problem-solving: thus any "artificial" intelligence would also be directed at problem-solving. That has set the tone for the goals of AI in all its multifarious applications. But intelligence is not only of the problem-solving variety; it also manifests itself in social and emotional contexts, for example. The tacit judgment required in those other contexts and exercised continuously by cognitively competent mature human beings has so far not been modeled in machines. Lacking what some logicians loosely call "intuition," it is not at all clear that AI can achieve the most ambitious (and perhaps perilous) milestones often attributed to it[1]. In the military setting, the outstanding question is whether AI can not only reliably contribute to strategic goals and tactical planning but is also effective at the operational level.

As this debate has evolved in the past decade, US defense planners have de-emphasized general AI and fully autonomous systems as a goal, perhaps partly in response to worries about a "doomsday device" with no human interruption possible, thus becoming too similar to an automatic weapon. In 2016, speaking of the US government's new doctrine for asymmetric advantage or "offset" over potential adversaries, Deputy Defense Secretary Robert Work remarked that "people say 'what's the Third Offset Strategy[2] about? And they say 'oh, it's about AI and autonomy.' We say no… It's about human-machine collaborative combat networks." While the reference to collaboration is reassuring, collaboration does not imply that absolute human control is always required. US Department of Defense directive 3000.09 on Autonomy in Weapon Systems currently requires that all systems "allow commanders and operators to exercise appropriate levels of human judgment over the use of force." In the absence of any system capable of general AI computer scientists focus on narrow AI, systems that can perform specific tasks for which they were trained, like the systems for playing complex games like chess and Go. The limits of narrow AI raise questions about hacking and other technical measures that may interfere with warfighter operations. Flaws in the algorithms that run narrow AI systems also raise ethical issues, as in the cases of racial and gender bias. Apart from an adversary's disruptive measures and biased coding, research and development of AI-enabled technology with warfighters itself poses ethical challenges that brain-computer interfaces (BCI) exemplify.

The emergence of AI-enabled brain-computer interfaces

BCI is a paradigmatic example of neurotechnology, understood as any technology that helps to influence and understand the brain and its functions. "A BCI is a computer-based system that acquires brain signals, analyzes them, and translates them into commands to an output device to carry out a desired action." Those signals are able to control cursors, prostheses, wheelchairs and other devices. "True" BCI systems use only signals from the central nervous system (CNS) and not from peripheral muscle nerves. In general, brain signal acquisition can be accomplished in two ways. Scalp-recorded EEG signals (eBCI) and wearable augmented realty (AR) systems are non-invasive (Portillo-Lara et al., 2021, p. 3). In contrast, intracortical microarrays (iBCI) vary from semi-invasive neural technologies, such as electrocorticography (ECoG) that require a craniotomy to place epidural or subdural electrodes on the surface of the cortex, to deeply embedded intracortical BCI or ocular or auditory implants.

These techniques have offsetting advantages and disadvantages. An eBCI is non-invasive but signal acquisition through the skull and scalp is difficult, whereas iBCI may improve signal strength but requires surgery and its attendant risks. Conventional improvements in BCI-based devices will function as therapeutic interventions, e.g., controlling prosthetics to restore capacity, including restoring nervous system feedback through artificial limbs. However, these devices can also maintain and enhance human performance during training and deployment. What is not settled, however, are the conditions under which these performance enhancements ought to be tested on or used by warfighters.

BCI predates AI by decades but can operate under GOFAI or the newer connectionist models. In the 1920's, the University of Jena's Hans Berger demonstrated the ability to read out electrical activity in the human brain *via* electroencephalography (EEG). The evolution of these fields illustrate how AI and BCI[3] have converged thanks to innovations in reading the brain's electrical

1 See for example Erik J. Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Cambridge, Mass: Harvard University Press, 2021).

2 For nonmilitary readers, "offsets" refer to the balance of force between nations, usually in great power conflict. The first offset in US doctrine is nuclear deterrence, and the second is stealth and precision guided munitions to counter larger hostile numerical forces. So, the "third offset" often gets used to describe some emerging set of technologies that will shift the balance of power, and usually (though not always) this is AI/autonomous systems.

3 For efficiency in this paper BCI will be assumed to include human computer interface (HCI) with specificity in the context of the discussion.

impulses[4]. In 1965, UCLA's Thelma Estrin articulated the requirements for a signal conversation system such that brain signals could be "digitized, filtered, classified and translated into cursor movement, for example, at very high speed." These were in effect the requirements for a BCI[5]. Also at UCLA, "direct brain-computer communication" was outlined by Vidal (1973). In the words of one history:

> "…the subject's EEG was to be transmitted to an amplifier the size of an entire desk belonging to the control area, which comprised two other screens and a printer. Then, after several steps, including analog-digital conversion, the signal would enter the IBM 360/91 for computing. Vidal asked, 'Can these observable electrical brain signals be put to work as carriers of information in man-computer communication or for the purpose of controlling such external apparatus as prosthetic devices or spaceships?' And he answered, 'Even on the sole basis of the present states of the art of computer science and neurophysiology, one may suggest that such a feat is potentially around the corner (Brunyé et al., 2014)."

In the 1970's and 1980's, it was noted that event-related potentials (ERPs) could be generated in response to external or internal stimuli. Biofeedback of EEG activity enabled subjects to engage in intentional activities like moving an image on a television screen or a cursor on a computer monitor. With "P" standing for "electrical positivity" and "300" for the delay in milliseconds between stimulation and voltage change, the so-called P300 wave allowed neurotypical volunteers to spell words on a computer screen (Shih et al., 2012). In the clinical setting, microelectrodes inserted into specific brain areas began to be experimentally employed in the early 2000's with patients suffering from loss of limb control. The case of spinal cord injury patient Matt Nagle was described in *Wired* in 2005. Nagle, who learned how to control a computer cursor, was a participant in a clinical tried called "BrainGate." Followed by BrainGate2, as reported in NRC (2009), brainstem stroke patient Cathy Hutchinson used a prosthetic arm to drink a bottle of coffee. These studies employed cables that tethered the patient-subject to brain signal-decoding computers, significantly limiting movements. In 2021, the BrainGate group announced successful experiments with an intracortical wireless BCI (an iBCI) with an external wireless transmitter.

Both the National Institutes of Health (NIH) and the Defense Advanced Research Projects Agency's (DARPA) Biological Technology Office (BTO) have committed to substantial investment in, *inter alia,* brain-computer interfaces connecting warfighters to computers through their brains. These neurotechnologies are a potential key to future US national defense, as well as a potential risk if developed by adversaries. More ambitious goals reach beyond simple EEG analysis and recording typical of implants and headsets to the use of AI to enhance BCI function is a central component of emerging military innovation. The BTO has described the ultimate goal of BCI as "BCI-AI fusion," where AI and a human user communicate bidirectionally to share control over a task or system. This combination of human and artificial cognition is seen as a key strategic asset in future conflicts. In launching the new BTO program "Next-Generation Nonsurgical Neurotechnology (N)," Almondi noted that "DARPA is preparing for a future in which a combination of unmanned systems, artificial intelligence, and cyber operations may cause conflicts to play out on timelines that are too short for humans to effectively manage with current technology alone." By connecting warfighters and decision makers to AI, rapid response to electronic and kinetic warfare can be managed using the skills humans and machines excel at, and keep a human in (or on) the loop in vital operations. In theory, the opportunities are remarkable. In the words of two IBM computer scientists, "[n]eurotech can interact with neurodata either invasively and directly through different kinds of surgical implants, like electrodes or devices implanted into or near neuronal tissues, or they can interact non-invasively and indirectly through wearable devices sitting on the surface of the skin…".

There is already high-level attention among military planners to these possibilities for technologically mediated cognitive enhancement, not all of which appear in the first instance to be relevant to AI. Commercial EEG-detection neurotechnologies in headsets like Emotiv and NeuroSky have garnered public attention but are not AI-enabled. However, military planners are anticipating the convergence of headsets and AI. In 2017 a US Navy Special Operations commander called for the development of a non-invasive brain stimulation (NIBS) device that uses electrical stimulation to improve performance. A product of the company Halo Neuroscience, the Halo Sport Headset (based on electrical stimulation *via* tDCS) was designed to improve physical performance but was noted anecdotally also to improve cognition. It is reported to have been tested on Navy SEALs at five sites for cognitive enhancement, resulting in improved performance, as in the case of ameliorating the consequences of sleep-deprivation. Although a NIBS device is not in itself AI-enabled, like many other neurotechnologies it can be linked to an AI system to record and modulate neural activity, potentially improving the efficacy of the enhancer. Such "closed loop" AI-enabled systems can self-correct using feedback

4 Elon Musk's Neuralink is the best known of these companies but there are other startups in this space with different approaches, especially in the ways that neural activity is recorded. Among the more innovative approaches, Stentrode introduces stents in blood vessels rather than some form of invasive bioelectrode or surface sensor.

5 Perhaps there was something in the water at that time in Los Angeles: Only a year later the first BCI that came to the attention of many Baby Boomers was the one featured in a 1966 Star Trek episode in which a severely brain injured Captain Christopher Pike uses such a system to communicate, but in this primitive approach the user was limited to one signal for "yes" and two signals for "no".

control to improve their devices' targeting and reliability. Nonetheless, if they modify cognition, even devices worn on the surface of the skin may be functionally equivalent to invasive devices.

Current state of military brain enhancement and ethics

Brain enhancement experiments (including BCI as a prominent example) have attracted notice in the US in the form of expert advisory reports. Here we note several of those produced mainly by the National Academies of Science, Engineering and Medicine (NASEM), as these are most relevant to warfighter enhancements and neurotechnologies. Several US presidential advisory commissions have also issued reports that are relevant more generally to experiments involving warfighters. Some consensus has crystallized around an intuitive definition of enhancement in terms of a contrast with therapeutic interventions. In their report *Beyond Therapy* (2003), the President's Council on Bioethics articulated that consensus view:

> "Therapy," on this view as in common understanding, is the use of biotechnical power to treat individuals with known diseases, disabilities, or impairments, in an attempt to restore them to a normal state of health and fitness. "Enhancement," by contrast, is the directed use of biotechnical power to alter, by direct intervention, not disease processes but the "normal" workings of the human body and psyche, to augment or improve their native capacities and performances[21].

Like the President's Council and other authorities, we find the distinction of enhancement versus therapy the most useful rule-of-thumb.

Of more immediate interest is the Council's concern that "biotechnical power" could be used to modify the human psyche in particular, well "beyond therapy," is what many find intuitively objectionable. Yet, as Lin et al. (2013) note in their research study on enhanced warfighters, "it is unclear how these objections would apply to the military context, e.g., whether they would be overcome by the special nature of military service and the exigencies of military operations…." Apart from the question of the acceptability of enhancement in the military setting in general, the acceptability of particular enhancements is a matter of perspective of different types of warfighters and their superiors, of their unit and third parties such as family members, of other military members, of civilians with whom they interact, of the government, and of the public and the nation. The history of modifying the human psyche "beyond therapy" is, moreover, arguably already common in many militaries in which the reluctance to kill other humans has been seen as a trait that needs to be trained out of warfighters (Evans and Hereth, forthcoming).

One of the few studies of its kind, the US National Academies report entitled *Opportunities in Neuroscience for Future Army Applications* (NRC, 2009) was an ambitious attempt to assess historical, ethical, and cultural issues for neuroscience in the army; neuropsychological testing in soldier selection, training, and learning; optimizing decision making; improving cognitive and behavioral performance ("hours of boredom and moments of terror"); neurotechnology opportunities like BCI; and long-term trends in research such as neural correlates for cultural differences in behavior. The same 2009 report described "in-helmet EEG for brain-machine interface" as a high-priority, medium-term (5–10-year) application opportunity. The report committee presciently emphasized that neither these kinds of opportunities, nor the points outlined in its 15 recommendations, would come to fruition without a single place in the Army to monitor potential neuroscience progress, evaluate potential applications and conduct the appropriate experimental research.

Perhaps surprisingly considering the subject of the report, although there is a section on the ethical issues raised by genetic screening of healthy persons, the report does not specifically address ethical issues about neurotechnologies beyond presupposing compliance with federal guidelines and regulations. It does raise the question of the applicability of research results derived from the usual volunteer subjects like undergraduate students, or even clinical patients, to a soldier population. Better surrogates might be high-performance athletes about whom there is extensive neuropsychological data. They may even be far superior subjects. When it comes to actual applications there are other challenges, including little knowledge of the candidate's psychology that may be relevant to their communication with other humans and to machines. In a chapter on neurotechnology opportunities, the report addresses issues like the physical load of any new device (not adding more than 1 kg to the helmet or 2 kg to the pack, not interfering with ballistic protection or helmet stability or freedom of head movement), field-deployable markers of neural state, EEG-based computer interfaces, haptic feedback for virtual reality, and augmented reality technologies, among others.

Ethical considerations for AI-enabled neurotechnology experimental research

Emerging AI-enabled neurotechnologies that may ultimately be operationally deployed present opportunities for warfighting and novel challenges to ethical standards for research and development involving warfighters. "Neuroenhancement" marries such life sciences as neurology, pharmacology, genetics, and psychology with long-time

soldiering attributes that include endurance, speed, intelligence-gathering, targeting, and training, none of which are medical conditions. As with any military technology, neuroenhancement products move slowly from research and development to field use.

At the research stage, ethical criteria require clinical investigators to establish the value and necessity of their proposed research, demonstrate a favorable cost/benefit ratio, utilize valid scientific methods, and protect research subjects' rights and welfare (Emanuel et al., 2000). Chief among research subjects' rights is informed consent that healthy volunteer research subjects must provide. Informed consent respects agents' dignity and right to self-determination by affording research subjects the information they require to weigh the costs and benefits of participating in medical research. Given the checkered history of military medical experimentation (Faden et al., 1995; Siegel-Itzkovich, 2009); however, the rules and regulations for clinical research among service personnel include special protections.

Following non-military clinical research protocols for vulnerable populations, military organizations in the US and Europe institute provisions to protect military research subjects' rights. Military officials understand that formal expressions of consent do not guarantee its respect. Although soldiers sign consent forms, problems arise because of rank disparity, fears of offending one's superiors, and/or peer pressure, which may undermine informed consent when soldiers are asked to participate in medical experiments (European Parliament, 2014, para. 31). As a result, additional regulations oversee clinical research and protect research subjects from coercion. The importance of voluntary consent is especially strong in cases where medical enhancements are *irreversible* (Davidovic and Crowell, 2022).

To safeguard voluntary consent among service members, The DoD's *Human Subjects Protection Regulatory Requirements* (Department of Defense, 2019, also: 32CFR219, "Protection of Human Subjects," and US Department of Defense Instruction 3216.02, 2018, 45 CFR 46, 2019) forbids the involvement of superior officers during the solicitation of research subjects and demands informed consent, medical supervision, the right to end an experiment, and the employment of an independent ombudsman or research monitor to oversee recruitment and experimentation [Department of Defense (DoD), 2011, p. 24–25]. British military officials, like their American counterparts, appoint an independent medical officer (IMO) to monitor the health, safety, and wellbeing of the participants (UK Ministry of Defense, 2020, p. 8; Linton, 2008).

To ensure that investigators meet statutory and ethical guidelines, independent and multidisciplinary Institutional Review Boards (IRB) in the United States (Department of Defense (DoD), 2011, p. 11–29), and Ministry of Defense Research Ethics Committees (MODREC) in the United Kingdom (UK Ministry of Defense, 2020), oversee research approval and compliance. Research oversight is complicated and time-consuming. Charged with what British officials term "proportionate scrutiny" (UK Ministry of Defense, 2020, para. 2–5), committee members seek a balance between outcomes and rights. Outcomes comprise benefits net of cost. Rights speak to respect for dignity and autonomous decision-making, informed consent, and acceptable risk.

These safeguards, however, are only part of the picture. They formally ensure informed consent, but researchers must provide adequate data to give substance to the right. Notice how emerging technologies pose medical risks for healthy research subjects while, at the same time, the operational goals of enhancement, that is, mission success, are entirely military. Therefore, ethically sustainable neuro-enhancement military research requires investigators to address two questions simultaneously so they may attain critical military goals while protecting research subjects' rights:

1. Is the proposed enhancement technology medically and militarily necessary?
2. Do the medical and military risks outweigh their benefits?

The following sections consider each of these questions in turn.

Medical necessity: What medical advantages does clinical research provide?

The overriding goal of any therapeutic clinical study is *medical* necessity. Investigators must demonstrate the likelihood that a new technology or medical procedure will not only effectively save lives or improve their quality but is also necessary. "Necessary" means that no other technology or procedure will attain the same outcome at a lower cost. There are no grounds to research a costly medical device, for example, if it is only as effective as a much less expensive existing technology. Therefore, it would be egregiously unethical to pursue unnecessary human research. However, non-therapeutic enhancements are neither curative nor rehabilitative. They do not save or improve the lives of the sick or injured. What medical benefit, then, do they provide warfighters? In what way are they *medically* necessary? One answer is that they are not. Enhancement provides research subjects with no medical benefits. Is conducting such research, therefore, ethically permissible?

There are two ways to address this objection. In one respect, enhancement research offers experimental subjects a personal benefit. As enhancement technologies push beyond normal baseline capabilities, they can boost a person's memory, sensory acuity, or targeting accuracy and, in this way, improve some

warfighters' chance of survival. While surviving one's occupation is immensely valuable to the survivor, it is nonetheless largely instrumental in a military context. By optimizing warfighter performance, successful enhancement improves the prospect of mission success. As it does, mission, not medical, success assumes the metric for measuring the necessity of cognitive enhancement research.

In saying this, we do not mean to assert that every warfighter enhancement *directly* benefits the enhanced individual. It probably does not. However, this leaves open the possibility that successful warfighter enhancements—i.e., enhancements that support strategic dominance and actualize military objectives—*indirectly* benefit enhanced individuals. As an analogy, consider vaccinations. As Jason Brennan observes,

> [T]he problem is that individuals as individuals make little difference. If everyone in the world were vaccinated except for Andy and Betty, Andy and Betty would pose no real threat to each other. Instead, vaccination presents a collective action problem, in which individuals as individuals are unimportant. [...] In general, individual decisions to vaccinate or not have negligible effects on others. What matters is what *most* people do, not what individuals do (Brennan, 2018, p. 39, 40).

When enough individuals are vaccinated, herd immunity is achieved. Herd immunity benefits *the herd*, a group of individuals, and by extension benefits *most members* of the herd. In a similar way, warfighter enhancements provide a kind of 'herd immunity' that protects against military failure, which in turn protects warfighters as a group and, therefore, *most individual* warfighters. Thus, the relevant kind of 'medical necessity' entailed by military necessity is equivalent to the kind of "medical necessity" entailed by public health necessity, as illustrated in the case of vaccinations.

Mission success, however, is fundamentally a *military*, not a *medical*, benefit that researchers and institutional review boards (IRBs) must weigh against a medical risk as they evaluate a project's feasibility. Like individual soldiers, IRBs face a utility calculation of incommensurable values: medical risks and military benefits. In practice, however, IRBs may resist this balancing act and instead search out individual medical or personal benefits, such as resiliency or language proficiency, that a research subject may acquire from participating in an experiment. But these personal advantages cannot be the determinative counterweight to individual risk in cognitive enhancement research. An enhancement technology that optimizes target selection, for example, may offer no discernable advantage to the research subject. In this situation, military benefits alone offset the medical risks of experimentation and provide the rationale for IRB ethics approval.

In this environment, researchers must proceed differently when conducting experimental studies than in clinical studies. They must convincingly argue that their proposed technology, a BCI, for example, is militarily necessary in the same way that therapeutic interventions are medically necessary. This requirement mirrors clinical guidelines that remind researchers, "because a normal healthy subject does not directly benefit from the study, the risk-benefit analysis must focus strongly on *the importance of the knowledge to be gained*" (e.g., Cornell University Office of Research Integrity, emphasis added). In this case, the knowledge gained is medical so that healthy research subjects must satisfy themselves that the greater good they serve (important medical knowledge) offsets the personal risk they incur during experimentation. In contrast, the critical knowledge provided by neuro-enhancement experimentation is primarily military. As a result, research subjects balance the medical risks of enhancement against its military benefits, a dramatically different sort of calculus to assess necessity.

Military necessity: What military advantages does enhancement research offer?

A recent RAND report (Binnendijk et al., 2020), *Brain Computer Interfaces: US Military Applications and Implications*, turns to military and technical specialists to evaluate brain-computer interfaces during urban operations in asymmetric war (p. 6). Using BCI as their test case, they asked: "which [BCI] capabilities [e.g., communication management, weapons control, enhancement cognitive or physical performance and training] were seen as more *useful* to support complex ground operations (emphasis added)." While the results certainly contribute to the BCI debate, the experimental design overlooks the question of necessity. Usefulness is not necessity. Asked to choose among seven BCI technologies, respondents were not asked to compare these to existing technologies that might improve training, weapons control, or communication. And while they may have been useful, there was no way to know if they were necessary and therefore, viable candidates for human research.

More critically, the RAND study's experimental design focused on a narrow range of counterinsurgency (COIN) operations: clearing a building of insurgents and evacuating wounded warfighters. This choice of cases raises two questions. First, how central are these tactical operations to asymmetric war? Second, is asymmetric war the paradigm we should use for evaluating BCI? One of us has argued, for example, that contemporary counterinsurgency warfare pushed well-beyond the kind of urban warfare described in the RAND report to include drone attacks, cyber and information warfare and, above all, population-centered counterinsurgency and public

diplomacy to win "hearts and minds (Gross, 2021, p. 181–203)." Among the neuro-enhanced skills required for COIN are language acquisition, cultural knowledge, and conflict management. The ideal soldier in modern asymmetric war may not be "a super-empowered soldier able to perform solo missions and transmit data back to headquarters" (Malet, 2015, p. 3); also (Galliott and Lotz, 2017), but one closer to Kaurin's description of a "Guardian." The Guardian embodies "soft" warfighting skills that attend to the needs of the weak and vulnerable, resolves issues without the use of force, pays attention to "culture, language and politics," and displays adaptability (Kaurin, 2014, p. 89–90).

Asymmetric war, moreover, is not the only game in town. On the one hand, NATO nations may intervene in conventional set-piece warfare as it currently wracks Ukraine. On the other, the West may veer toward near-peer confrontations with China or Russia or confront nuclear threats from Iran and North Korea. In the latter instances, emphasis shifts from traditional warfighting concerns of offsetting troop strength and military assets to offsetting an adversary's rapid technological advancements. New technologies include advanced computing, "big data" analytics, artificial intelligence, autonomy, robotics, directed energy, hypersonics, and biotechnology [Department of Defense (DoD), 2018]. In the words of one group of Chinese neuroscientists, "Artificial intelligence (AI), which can advance the analysis and decoding of neural activity, has turbocharged the field of BCI" (Zhang et al., 2020).

With the "turbocharging" of BCI by AI in mind and considering the scenarios of contemporary and near-term warfare one must ask where and how neurotechnologies like BCI are useful and necessary in these contexts. What is this technology's highest and best use? While implantable iBCI may enable a generation of bionic warfighters (Britzky, 2019), their role in contemporary and future warfare remains unsubstantiated and, perhaps, marginal. In contrast, EEG-based eBCI significantly improve training and learning by offering feedback loops to evaluate data and monitor performance by a human operator. Similarly, non-invasive nerve stimulation devices such as earbud electrodes enable targeted neuroplasticity training (TNT) to accelerate language acquisition, acculturation, and intelligence analysis to facilitate successful population-centered COIN. eBCI and other TNT neuro-technologies help operators organize information flows to permit fast-moving threat and target identification (Naufel et al., 2020). In these ways, eBCI do not enhance the killing capabilities that some iBCI may offer warfighters. Instead, they can improve the quality of the intelligence warfighters receive while enhancing the soft skills required to attend to the needs of the local population.

Evaluating military necessity at the research stage is a speculative but essential endeavor that should integrate military analysts into the preparation of clinical studies. But the absence of any sustained discussion of military necessity is glaring. Nevertheless, many researchers avoid the discussion of military benefits altogether or only offer perfunctory details. A 2019 consent form from the US Army Aeromedical Research Laboratory, for example, makes short shrift of potential military benefits of anti-fatigue agents. It simply advises potential research subjects, "*Your participation will contribute to the medical knowledge and scientific investigation of possible uses for these medications in a military operational setting.*" Under UK Ministry of Defense Research Ethics Committee (MODREC) guidelines entitled, "Participant Involvement: Risks, Requirements and Benefits," Paragraph 17h instructs researchers to "*describe any expected benefits to the research participant (if none, state none).*" "None" only makes sense if the expected benefits are solely medical. In neither example do researchers "focus strongly on the knowledge to be gained" from experimentation. To do so will inevitably draw military policymakers and ethicists into enhancement research.

To provide fully informed and voluntary consent, research subjects must also contend with military and medical *risks*. Medical risks may be physiological and/or psychological and may render some technologies that require surgical implantation, for example, unsustainable. Here, issues related to the vulnerability of specific populations come into play. Military risk is both technological and organizational. The former includes vulnerability to hacking and data theft, while the latter raises concerns about disseminating and protecting data among the many interested stakeholders in a military organization.

Medical risks

Surgically implanted brain-computer interfaces pose significant medical risks leading DARPA to reject surgically invasive enhancement techniques:

> Due to the inherent risks of surgery, these technologies have so far been limited to use by volunteers with clinical need. For the military's prima-rily able- bodied population to benefit from neurotechnology, *non-surgical interfaces are required.* Teams are pursuing a range of approaches that use optics, acoustics, and electromagnetics to record neural activity and/ or send signals back to the brain at high speed and resolution. The re-search is split between two tracks. Teams are pursuing either completely non- invasive interfaces that are entirely external to the body or minutely invasive interface systems that include nanotransducers that can be tempo-rarily and non-surgically delivered to the brain to improve signal resolution [Defense Advanced Research Projects Agency (DARPA)., 2019, emphasis added].

Some observers concur: "To effectively implement BCI systems… for enabling *efficient performance by healthy users*," write Miranda et al. (2015, p. 64), "there exists a need for the development of subcutaneous and *fully non-invasive* neural interfaces that are both portable and capable of recording activity

from cortical and deep brain structures at high spatial and temporal resolution (emphasis added)." However, others draw a line between research and deployment. "Despite the high accuracy and optimal signal fidelity [of intracortical electrodes]," write Portillo-Lara et al. (2021, p. 3), "the risks associated with the surgical procedures largely restrict their use outside well-controlled laboratory and clinical environments." Similarly, "greater risk may be tolerable for the restorative technologies… in the clinical domains, but could be less ethically justifiable for the performance benefits for healthy individuals" (Naufel and Klein, 2020, p. 5).

Rejections of high-risk, implantable neurotechnologies for healthy individuals are *de rigueur* but not always accompanied by convincing ethical arguments. Despite legitimate apprehension about coercion and undue influence that comes from "institutional or hierarchical dependency (European Parliament, 2014, para. 31)," military personnel are not a vulnerable population on par with minors, prisoners, or the economically disadvantaged, as some suggest (McManus et al., 2007; Parasidis, 2016). Service personnel do not lack sound decision-making capacity or suffer from socially inflicted disabilities. There are no a priori reasons that render service personnel incapable of making informed choices about their participation in medical research or willingness to accept these risks if counterbalanced by military or, to a lesser extent, medical benefits.

Researchers may also reject invasive neuroenhancements because they believe the risk is too high or insufficiently known (e.g., Nijboer et al., 2013, p. 553). Naufel and Klein (2020, p. 2) cite a 20–40% risk of surgical complications and 24–50% risk of hardware complications. Additionally, researchers and funding agencies may think alternative semi-invasive or non-invasive neurotechnologies are adequate for military purposes. Whether implantable technologies are *necessary* is a logically prior question that demands an answer before considering surgical risks. Until it is, there is no *prima facie* reason to reject invasive technologies.

bib44 If implantable BCI pose the danger of surgery and interface maintenance, eBCI are not entirely without risk. Researchers note unknown psychological risks affecting personality, memory, and BCI dependence (Vlek et al., 2012; Kögel et al., 2019; National Academies of Sciences Engineering Medicine, 2021, p. 41, 50). Incorporating AI in BCI adds additional unpredictability and risk. Unlike traditional BCI, whose functions may be static, a self-correcting AI can dynamically adapt how it operates. As a result, additional risks may accumulate as research subjects interact with BCI and AI-enabled BCI react and adapt to stimuli.

Nevertheless, evaluating such risks is integral to the research project. As such, research subjects require a good-faith assessment of these risks and the means to mitigate them should adverse psychological effects or unpredicted AI adaptations surface during or after the experiment. It is challenging to present potential psychological or AI-related risks to research subjects when their full extent is unknown until the trial concludes. Phase 1 drug trials, for example, investigate toxicity. As such, research subjects cannot receive but scant information about potential risks. However, buoyed by optimism and "therapeutic misestimation" that exaggerates a trial's benefits, critically ill research subjects often discount the risks and consent to experimental treatment (Miller and Joffe, 2013; Halpern et al., 2019). However, military research subjects for cognitive enhancement are not ill. There are few or no medical benefits to excite sufficient sanguinity to offset thinly demonstrable risks. As a result, non-therapeutic researchers operate under stricter conditions than clinical researchers. We can only speculate about the psychological effects of BCI (personality changes, memory disruptions, or BCI dependence) and the additional risks of AI-enabled BCI because one research goal is to study these effects. But to obtain fully informed and voluntary consent, research subjects also need additional data about technological and institutional risk.

Risks: Technological and institutional risks to privacy and confidentiality

Technological risks comprise BCI hacking that may put personal information in hostile hands. Institutional risks come when myriad stakeholders claim privileged information, including related agencies, the scientific community, pharmaceutical companies, and perhaps, allied nations. This coterie of stakeholders is not unique in military medicine, where patients have limited rights to their personal medical data (Gross, 2021). Technological and institutional risks impinge upon privacy and confidentiality, two fundamental rights of research subjects.

Privacy and confidentiality are closely related. Privacy is a subsidiary right of personal self-determination: the right to keep information close and release only what one wants others to know about oneself (Bok, 1989, p. 120). Confidentiality is a duty imposed on others to guard another's private information until that person authorizes its disclosure. The right to privacy and the duty of confidentiality ensure self-esteem, job security, and social status that the release of personal information may jeopardize. In medicine, respect for privacy preserves the trust necessary for practitioners to tend patients successfully and for researchers to maintain the trust they need to conduct clinical trials. Usually, privacy and confidentiality are straightforward. Patients disclose information so medical practitioners can provide proper care. Beyond that, it is nobody's business.

Novel risks to autonomy are also raised by the prospect of neurointerventions. For example, deep-brain stimulation (DBS) applied therapeutically to Parkinson's patients has undermined patients' sense of personal authenticity and enhanced their sense of alienation, leading some (e.g., Kraemer, 2013) to conclude that DBS poses serious risks for autonomy, and

others to propose non-individualistic conceptions of autonomy (Lee, 2021). Indeed, some scholars contend that theoretical neurointerventions provide a basis for ethical theorizing about the nature of autonomy (Zuk and Lázaro-Muñoz, 2021). By contrast, other scholars like Douglas (2022) argue that just as "nudges" can treat their targets as rational agents, so too can non-consensual neurointerventions. Plausibly, the possibility of treating one's targets as rational agents entails the possible retention of their autonomy, such that even *non-consensual* neurointerventions might respect autonomy (cf. Gillett, 2009). Even more controversial is Pugh (2014) claim that some neurointerventions, such as those that reduce impulsivity, can *enhance* patient autonomy (cf. Fleishmann and Kaliski, 2017).

Clinical research is bound by weaker rules of privacy than medical practice is. For example, research subjects may be required to share large chunks of anonymized data as part of the experimental research (Malin et al., 2010). In addition, AI-assisted enhancement research may further attenuate privacy, thereby requiring researchers to provide healthy research subjects with answers to the following questions:

1. Data attributes: What kind of data and in what format do BCI record? What personal or ancillary information do the data reveal?
2. Data accessibility and sharing: Who has access to the data? What agreements are there for data sharing? Who can potentially read this data?
3. Data protection: How are the data protected? Where are the data stored during and after the experiment? Are the experimental BCI vulnerable to hacking as some fear?

The answers to these questions are the subject of research itself. Most iBCI use intracortical devices to measure neuron activation potential in particular brain regions, often on the level of individual neurons. eBCI tends to use fMRI or EEG signals. Both signals measure activation potential, usually across large segments or the whole of the brain. Typically, voltages or activation potentials correspond to particular mental states. These are neural correlates the machine receives as the basis for action. As a result, there are concerns regarding invasions of privacy, unauthorized access to confidential information and hacking. In response, data management plans, software fault tree testing, and red teams (that try to hack the machine on behalf of the manufacturer) address these concerns. They are integral to a research ethics protocol (Denning et al., 2009).

Finally, while the technological risks associated with utilizing AI are broad and cannot be adequately summarized in this paper, we would be remiss if we failed to mention a few crucial areas of concern. First, AI has well-known racial (Kostick-Quenet et al., 2022)[6], gender (Wellner and Rothman, 2020; Waelen and Wieczorek, 2022), and disability biases (Tilmes, 2022). These algorithmic biases undermine the permissibility of unthinking reliance on purportedly "unbiased" AI. Second, AI decision-making is notoriously opaque – its decisions are made, as multiple scholars have described it, in an "algorithmic black box" (Hollanek, 2020; von Eschenbach, 2021). Despite occasional optimism about rendering AI decision-making transparent (e.g., Mishra, 2021), most scholars remain concerned about the effects of biased AI used for medical purposes. Among these are concerns that biased AI will reduce persons to mere data (Sparrow and Hatherley, 2019), that AI might impermissibly (and invisibly) incorporate economic data in its rationing recommendations (Sparrow and Hatherley, 2020; Braun et al., 2021), and that AI will rely upon other value-laden considerations (Ratti and Graves, 2022). Again, this is merely a sampling of the technological risks associated with AI. The risks extend well-beyond algorithmic bias. Yet these risks must be considered when evaluating the permissibility of AI-enabled warfighter enhancements.

Moving forward: Sustainable research ethics for neuroenhancement military research

Research protocols for *therapeutic* neurotechnologies draw attention to respect for autonomy, informed consent and self-determination, the right to privacy and confidentiality, and constant concern for the welfare of subjects, their community, and end-users (Girling et al., 2017; Pham et al., 2018). To maintain the same respect for the rights of healthy research subjects who participate in *non-therapeutic* military neuroenhancement research demands attention to a full array of unique military and medical costs and benefits. Therefore, any sustainable ethics protocol for non-therapeutic neuro-enhancement military research must closely note military and medical risks and benefits to adequately protect research subjects' rights. To date, most researchers fail to fully account for a novel technology's expected military benefits, sometimes over-compensate for military research subjects' vulnerability, fail to consider the technological and institutional risks to privacy and confidentiality and overlook the intricacy of balancing often incommensurable apples (medical risks) and oranges (military necessity).

Research subjects, therefore, adopt a utility calculus common in the military that positions personal risk against collective benefits. By taking stock of national or military interests, they may accept considerable personal risk if the military benefits accruing to their political commonwealth are significant. Attention to military necessity and collective social interests at the expense of individual wellbeing is not foreign to military medical ethics. The US Army Medical Department (AMEDD) *Emergency War Surgery* (Cubano, 2018), for example, reminds its per-sonnel: "the ultimate goals of

6 Interestingly, the use of AI – in particular, the use of avatars – can *reduce* implicit racial bias (Peck et al., 2013). Thus, the use of even racially biased AI could theoretically mitigate racial biases in human users.

combat medicine are the return of the greatest possible number of warfighters to combat and the preservation of life, limb, and eyesight. Commitment of resources should be decided first based on the mission and immediate tactical situation *and then* by medical necessity, irrespective of a casualty's national or combatant status" (Cubano, 2018, p. 24, emphasis added; cf. JP 4- 02, 2001: II- 1; 2006:ix). And while this provision applies to therapeutic care, it informs research priorities as well.

More data and greater sensitivity drive the way forward. AI-enabled neuroenhancement offers tremendous possibilities for military use to improve warfighting capabilities, reduce service members' exposure to life-threatening danger and meet emerging threats. But sensitive to research subjects' rights, investigators must spell out the military advantages in far greater detail while IRBs supervise compliance. Although data collected from large numbers of healthy, young warfighters may turn out to be instructive for medical science, no military medical research protocol should content itself with simply telling subjects that they are taking significant risks for *medical knowledge . . . in a military operational setting*. Non-therapeutic military neuro-enhancement research protocols also cannot suffice with compiling medical risks alone. Moreover, there are ethically relevant differences between clinical research and non-therapeutic military medical research which draws in vested stakeholders and parties with access to information. Unlike clinical medical research, military medical research is likely to attract hostile parties who may put subjects at considerable risk. In this way, neuro-enhanced soldiers share the attributes of newly developed weapons, and their nations must acknowledge the danger they face and protect them accordingly.

Despite two decades of speculation about the prospects for neuro-enhancement amid the convergence of BCI and AI, an array of ethical issues that remain to be sorted out have been an obstacle to the systematic investigation of operational potential. To fill the lacunae of basic BCI/AI research, we have suggested a comprehensive and critical analysis of military necessity comparable to medical necessity. Medical necessity recounts the overwhelming advantage a new technology, intervention, or drug will offer individual patients and society. Military necessity must do the same for neurotechnologies designed to enhance warfighter performance while taking account of the conditions necessary to obtain fully informed and voluntary consent.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This paper was funded under a grant by the U.S. Air Force Office of Scientific Research, award number FA9550-21–1-0142.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Australian DoD (Department of Defence) (2020). *A Method for Ethical AI in Defence*. Defence Science and Technology Group, DSTG-TR-3786. Canberra: Aerospace Division, Defence Science and Technology Group.

Binnendijk, A., Marler, T., and Bartels, E. M. (2020). "*Brain-Computer Interfaces US Military Applications and Implications, An Initial Assessment," Rand Corp*. Available online at: https://www.rand.org/pubs/research_reports/RR2996.html (accessed August 16, 2022).

Bok, S. (1989). *Secrets: On the Ethics of Concealment and Revelation*. New York, NY: Vintage.

Braun, M., Hummel, P., Beck, S., and Dabrock, P. (2021). Primer on an ethics of AI-based decision support systems in the clinic. *J. Med. Ethics* 47, 3–3. doi: 10.1136/medethics-2019-105860

Brennan, J. (2018). A libertarian case for mandatory vaccination. *J. Med. Ethics* 44, 37–43. doi: 10.1136/medethics-2016-103486

Britzky, H. (2019). *The Army Wants to Stick Cyborg Implants into Soldiers by 2050 and it's Absolutely Insane. Task and Purpose*. Available online at: https://taskandpurpose.com/news/army-cyborg-soldier-2050-study/ (accessed July 26, 2022).

Brunyé, T. T., Holmes, A., Cantelon, J., et al. (2014). Direct current brain stimulation enhances navigation efficiency in individuals with low spatial sense of direction. *Neuroreport* 25, 1175–1179. doi: 10.1097/WNR.0000000000000214

Cornell University Office of Research Integrity and Assurance Human Research Participant Protection Program Sop 13: Informed Consent, Enrollment, and Other Considerations for Research Involving Normal, Healthy Participants. Available online at: https://researchservices.cornell.edu/sites/default/files/2019-05/

SOP%2013%20-%20Normal%20Healthy%20Volunteers.pdf (accessed August 16, 2022).

Cubano, M. A. (2018). *Emergency War Surgery, 5th revision* (Falls Church, VA: Office of The Surgeon General).

Davidovic, J., and Crowell, F. S. (2022). Operationalizing the ethics of solider enhancement. *J. Mil. Ethics* 20, 180–199. doi: 10.1080/15027570.2021.2018176

Defense Advanced Research Projects Agency (DARPA). (2019). *Six Paths to the Nonsurgical Future of Brain-Machine Interfaces. DARPA News and Events*. Available online at: https://www.darpa.mil/news-events/2019-05-20 (accessed August 16, 2022).

Denning, T., Matsuoka, Y., and Kohno, T. (2009). Neurosecurity: security and privacy for neural devices. *Neurosurg. Focus* 27, E7. doi: 10.3171/2009.4.FOCUS0985

Department of Defense (DoD) (2011). *Instruction Number 3216.02, November 8, 2011: Protection of Human Subjects and Adherence to Ethical Standards in DoD-Supported Research, Enclosure 3, Paragraph 9: Unique DoD Limitations on Waiver of Informed Consent (US)*. Available online at: http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/321602p.pdf (accessed August 16, 2022).

Department of Defense (DoD) (2018). *Summary of the 2018 National Defense Strategy of the United States: Sharpening the American Military's Competitive Edge*. Available online at: https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf (accessed August 16, 2022).

Department of Defense. (2019). *Human Subjects Protection Regulatory*. Available online at: https://mrdc.amedd.army.mil/assets/docs/orp/HRPO_Information_for_Investigators.docx

Douglas, T. (2022). If nudges treat their targets as rational agents, nonconsensual neurointerventions can too. *Ethical Theory Moral Pract.* 25, 369–384. doi: 10.1007/s10677-022-10285-w

Emanuel, E. J., Wendler, D., and Grady, C. (2000). What makes clinical research ethical?. *JAMA*. 283(20), 2701–2711. doi: 10.1001/jama.283.20.2701

European Parliament (2014). EU. *Regulation (EU) No 536/ 2014 on Clinical Trials on Medicinal Products for Human Use, and Repealing Directive 2001/20/ EC*. Available online at: https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-1/reg_2014_536/reg_2014_536_en.pdf (accessed August 16, 2022).

Evans, N., and Hereth B. (forthcoming). Can we justify military enhancements? Some yes, most no. *Camb. Q. Healthc. Ethics*.

Faden, R. R., Feinberg, K. R., Olenick, N. L., Glatstein, E., Royal, H. D., Katz, J., et al. (1995). *Final Report of the Advisory Committee on Human Radiation Experiments*. U. S. Government Printing Office. Available online at: https://www.osti.gov/opennet/servlets/purl/120931/120931.pdf

Fleishmann, A., and Kaliski, A. (2017). Personal autonomy and authenticity: adolescents' discretionary use of methylphenidate. *Neuroethics* 10, 419–430. doi: 10.1007/s12152-017-9338-3

Galliott, J., and Lotz, M. (2017). "Introduction," In *Super Soldiers: The Ethical, Legal and Social Implications*, eds. J. Galliott and M. Lotz (London: Routledge), 1–8.

Gillett, G. (2009). Intention, autonomy, and brain events. *Bioethics* 23, 330–339. doi: 10.1111/j.1467-8519.2009.01726.x

Girling, K., Thorpe, J., and Auger, A. (2017). Identifying ethical issues of human enhancement technologies in the military defence research and development canada, scientific report, DRDC-RDDC-2017-R103 October 2017.

Gross, M. L. (2021). *Military Medical Ethics in Contemporary Armed Conflict: Mobilizing Medicine in the Pursuit of Just War*. Oxford University Press.

Halpern, J., Paolo, D., and Huang, A. (2019). Informed consent for early-phase clinical trials: therapeutic misestimation, unrealistic optimism and appreciation. *J. Med. Ethics* 45, 384–387. doi: 10.1136/medethics-2018-105226

Hereth, B. (2022). Moral neuroenhancements for prisoners of war. *Neuroethics* 15, 1–20. doi: 10.1007/s12152-022-09482-2

Hollanek, T. (2020). AI transparency: a matter of reconciling design with critique. *AI and Society*. 1–9. doi: 10.1007/s00146-020-01110-y. [Epub ahead of print].

ICRC (2006). A guide to the legal review of new weapons, means and methods of warfare: measures to implement article 36 of additional protocol I of 1977. *IRCC* 88, 931–956. doi: 10.1017/S1816383107000938

Kaurin, P. M. (2014). *The Warrior, Military Ethics and Contemporary Warfare: Achilles Goes Asymmetrical* (London: Routledge).

Kögel, J., Schmid, J. R., Jox, R. J., and Friedrich, O. (2019). Using brain-computer interfaces: a scoping review of studies employing social research methods. *BMC Med. Ethics* 20, 1–17. doi: 10.1186/s12910-019-0354-1

Kostick-Quenet, K., Cohen, I. G., Gerke, S., et al. (2022). Mitigating bias in machine learning. *J. Law Med. Ethics* 50, 92–100. doi: 10.1017/jme.2022.13

Kraemer, F. (2013). My, myself, and my brain implant: deep brain stimulation raises questions of personal authenticity and alienation. *Neuroethics* 6, 483–497. doi: 10.1007/s12152-011-9115-7

Lee, J. Y. (2021). Revisiting moral bioenhancement and autonomy. *Neuroethics* 14, 529–539. doi: 10.1007/s12152-021-09470-y

Lin, P., Mehlman, M., and Abney, K. (2013). *Enhanced Warfighters: Risk, Ethics and Policy*. Available online at: https://case.edu/law/sites/case.edu.law/files/

Linton, R. (2008). Applying for ethical approval from the MoD research ethics committee. *J. R. Nav. Med. Serv.* 94, 41–46. doi: 10.1136/jrnms-94-41

Malet, D. (2015). Captain America in international relations: the biotech revolution in military affairs. *Def. Stud.* 15, 1–21. doi: 10.1080/14702436.2015.1113665

Malin, B., Karp, D., and Scheuermann, R. H. (2010). Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J. Investig. Med.* 58, 1–17. doi: 10.2310/JIM.0b013e3181c9b2ea

McManus, J., McClinton, A., Gerhardt, R., and Morris, M. (2007). Performance of ethical military research is possible: on and off the battlefield. *Sci. Eng. Ethics* 13, 297–303.? doi: 10.1007/s11948-007-9022-9

Miller, F. G., and Joffe, S. (2013). Phase 1 oncology trials and informed consent. *J. Med. Ethics* 39, 761–764. doi: 10.1136/medethics-2012-100832

Miranda, R. A., Casebeer, W. D., Hein, A. M., Judy, J. W., Krotkov, E. P., Laabs, T. L., et al. (2015). DARPA-funded efforts in the development of novel brain–computer interface technologies. *J. Neurosci. Methods* 244, 52–67. doi: 10.1016/j.jneumeth.2014.07.019

Mishra, A. (2021). Transparent AI: reliabilist and proud. *J. Med. Ethics* 47, 341–342. doi: 10.1136/medethics-2021-107352

National Academies of Sciences Engineering and Medicine (2021). *Human-AI Teaming: State of the Art and Research Needs* (Washington, DC: The National Academies Press).

Naufel, S., and Klein, E. (2020). Brain–computer interface (BCI) researcher, perspectives on neural data ownership and privacy. *J. Neural Eng.* 17, 016039. doi: 10.1088/1741-2552/ab5b7f

Naufel, S., Knaack, G. L., Miranda, R., Best, T. K., Fitzpatrick, K., Emondi, A. A., et al. (2020). DARPA investment in peripheral nerve interfaces for prosthetics, prescriptions, and plasticity. *J. Neurosci. Methods* 332, 108539. doi: 10.1016/j.jneumeth.2019.108539

Nijboer, F., Clausen, J., Allison, B. Z., and Haselager, P. (2013). The asilomar survey: Stakeholders' opinions on ethical issues related to brain-computer interfacing. *Neuroethics* 6, 541–578. doi: 10.1007/s12152-011-9132-6

NRC. (2009). *Committee on Opportunities in Neuroscience for Future Army Applications. Board on Army Science and Technology Division on Engineering and Physical Sciences*. National Research Council. Washington, DC: The National Academies Press.

Parasidis, E. (2016). The military biomedical complex: are service members a vulnerable population. *Houst. J. Health Law Policy* 16, 113–161.

Peck, T. C., Seinfeld, S., Aglioti, S. M., and Slater, M. (2013). Putting yourself in the skin of a black avatar reduces implicit racial bias. *Conscious. Cogn.* 22, 779–787. doi: 10.1016/j.concog.2013.04.016

Pham, M., Goering, S., Sample, M., Huggins, J. E., and Klein, E. (2018). Asilomar survey: researcher perspectives on ethical principles and guidelines for BCI research. *Brain-Comput. Interfaces* 5, 97–111. doi: 10.1080/2326263X.2018.1530010

Portillo-Lara, R., Tahirbegi, B., Chapman, C. A. R., et al. (2021). Mind the gap: state-of-the-art technologies and applications for EEG-based brain–computer interfaces. *APL Bioeng.* 5, 031507. doi: 10.1063/5.0047237

Pugh, J. (2014). Enhancing autonomy by reducing impulsivity: the case of ADHD. *Neuroethics* 7, 373–375. doi: 10.1007/s12152-014-9202-7

Ratti, E., and Graves, M. (2022). Explainable machine learning practices: opening another black box for reliable medical AI. *AI Ethics*. 1–14. doi: 10.1007/s43681-022-00141-z

Shih, J. J., Krusienski, D. J., and Wolpaw, J. R. (2012). "Brain-Computer Interfaces in Medicine," in *Mayo Clinic Proceedings*. 87, 268–279.

Siegel-Itzkovich, J. (2009). IDF's anthrax vaccine trial violated Helsinki Convention. *BMJ*. 338, b1325–b1325. doi: 10.1136/bmj.b1325

Sparrow, R., and Hatherley, J. J. (2019). The promise and perils of AI in medicine. *Int. J. Chin. Comp. Philos. Med.* 17, 79–109. doi: 10.24112/ijccpm.171678

Sparrow, R., and Hatherley, J. J. (2020). High hopes for deep medicine? AI, economics, and the future of care. *Hastings Cent Rep*. 50, 14–17. doi: 10.1002/hast.1079

Tilmes, N. (2022). Disability, fairness, and algorithmic bias in AI recruitment. *Ethics Inf. Technol.* 24, 1–13. doi: 10.1007/s10676-022-09633-2

UK Ministry of Defence (2021). *Human Augmentation – The Dawn of a New* Paradigm, A strategic implications project.

UK Ministry of Defense (2020). *JSP 536. Governance of Research Involving Human, Participants, Part 1: Directive*. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872936/20200312-JSP536_Part_1_Governance_Research_Human_v3_1_FINAL.pdf (accessed August 16, 2022).

US Department of Defense Instruction 3216.02. (2018). Protection of human subjects and adherence to ethical standards in DoD-supported research.

Vidal, J. J. (1973). Toward direct brain-computer communication. *Annu. Rev. Biophys. Bioeng.* 2, 157–180. doi: 10.1146/annurev.bb.02.060173.001105

Vlek, R. J., Steines, D., Szibbo, D., et al. (2012). Ethical issues in brain–computer interface research, development, and dissemination. *J. Neurol. Phys. Ther*. 36, 94–99. doi: 10.1097/NPT.0b013e31825064cc

von Eschenbach, W. J. (2021). Transparency and the black box problem: why we do not trust AI. *Philos. Technol.* 34, 1607–1622. doi: 10.1007/s13347-021-00477-0

Waelen, R., and Wieczorek, M. (2022). The struggle for AI's recognition: understanding the normative implications of gender bias in AI with Honneth's theory of recognition. *Philos. Technol.* 35, 1–17. doi: 10.1007/s13347-022-00548-w

Wellner, G., and Rothman, T. (2020). Feminist AI: can we expect our AI systems to become feminist? *Philos. Technol.* 33, 191–205. doi: 10.1007/s13347-019-00352-z

Zhang, X., Ma, Z., Zheng, H., Li, T., Chen, K., Wang, X., et al. (2020) The combination of brain-computer interfaces and artificial intelligence: applications and challenges. *Ann Transl Med*. 8:712. doi: 10.21037/atm.2019.11.109

Zuk, P., and Lázaro-Muñoz, G. (2021). DBS and autonomy: clarifying the role of theoretical neuroethics. *Neuroethics* 14, 83–93. doi: 10.1007/s12152-019-09417-4

frontiers | Frontiers in Big Data

TYPE Perspective
PUBLISHED 12 September 2022
DOI 10.3389/fdata.2022.991759

OPEN ACCESS

EDITED BY
Kirsi Helkala,
Norwegian Defence University College, Norway

REVIEWED BY
Grethe Østby,
Norwegian University of Science and Technology, Norway
James Cook,
United States Air Force Academy, United States

*CORRESPONDENCE
Neil C. Rowe
ncrowe@nps.edu

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 11 July 2022
ACCEPTED 19 August 2022
PUBLISHED 12 September 2022

CITATION
Rowe NC (2022) The comparative ethics of artificial-intelligence methods for military applications. *Front. Big Data* 5:991759. doi: 10.3389/fdata.2022.991759

The comparative ethics of artificial-intelligence methods for military applications

Neil C. Rowe*

Department of Computer Science, U.S. Naval Postgraduate School, Monterey, CA, United States

Concerns about the ethics of the use of artificial intelligence by militaries have insufficiently addressed the differences between the methods (algorithms) that such software provides. These methods are discussed and key differences are identified that affect their ethical military use, most notably for lethal autonomous systems. Possible mitigations of ethical problems are discussed such as sharing decision-making with humans, better testing of the software, providing explanations of what is being done, looking for biases, and putting explicit ethics into the software. The best mitigation in many cases is explaining reasoning and calculations to aid transparency.

KEYWORDS

artificial intelligence, ethics, algorithms, military, lethal autonomous systems, explanation, transparency

Introduction

Artificial-intelligence (AI) software is increasingly proposed to replace humans in military technology and military planning involving potential lethal force. Military conflict is dangerous, and there is much incentive to automate its actors. For instance, an automated gun turret from South Korea that uses simple AI is internationally popular (Parkin, 2015) although its ethical principles have not been carefully evaluated. The most obvious ethical issues with military AI occur with targeting, and other issues arise in the planning of operations and logistics support. However, building AI systems to make potentially lethal judgments is difficult, and current AI methods are still less accurate than humans for many tasks (Emery, 2021). Using them to apply lethal force can be unethical, just as using an imprecise weapon like a shotgun in military conflict today. Furthermore, a major justification for the use of lethal force in the laws of armed conflict is self-defense, something less relevant to software and robots since they can be cheaply remanufactured, although limited self-defense could still be appropriate for them to preserve their capabilities during an ongoing conflict. So it is important to assess how each AI method works to see how well its contribution to lethal force can be justified, and the methods differ considerably in their accuracy and explainability, and hence their possible justifiability.

Not all ethical problems of AI systems can be blamed on the software, as problems can be due to errors in input data, misconfiguration of systems, or deliberate sabotage. Furthermore, many software problems of AI systems cannot be blamed on AI, since AI depends on man-machine interfaces, databases, and networking that can also be faulty. Also, successful development of AI, like that of other software, depends on familiarity with the context in which it will be used, and AI developers rarely have experience in warfare that they can use in developing military AI systems. We do not consider those problems here as we wish to focus exclusively on problems of AI methods in this short article. Note that AI methods are not necessarily less ethical than those of humans, since such methods can exceed human capabilities in speed, accuracy, and reliability when designed, debugged, and tested well, and this could enable them make ethical decisions more accurately than humans in challenging situations.

Artificial-intelligence methods

Artificial intelligence is generally considered to be methods for creating intelligent behavior with software, not necessarily human methods. It is a form of automation, and raises some of the same issues as other kinds of automation, plus a few new ones because of its focus on information rather than machinery. Most ethical theories ascribe blame for unethical actions of algorithms (methods) to their creators and deployers; (Tsamado et al., 2021) identifies a wide range of possible ethical issues with algorithms. AI algorithms fall into several categories as described in (Rowe, 2022):

- Logical inference methods that reach yes/no conclusions assuming a set of starting facts. This includes rule-based systems with if-then rules, decision trees, and reasoning by analogy.
- Uncertain inference methods that reach conclusions with an associated degree of certainty, assuming a set of starting assertions with probabilities. This includes most artificial neural networks, Bayesian reasoning, and case-based nearest-neighbors reasoning.
- Planning and search methods that find good sequences or plans to solve problems using logical reasoning. This includes heuristic depth-first search, heuristic breadth-first search, and hierarchical planning.
- Planning and search methods that find good sequences or plans with an associated degree of estimated quality or certainty. This includes A* search, game search, and recurrent neural networks.
- Machine learning (the usual name for learning methods in the AI field) and other optimization of logical inference methods from examples of desired behavior. This includes set-covering methods, decision trees, and support-vector machines.
- Machine learning and optimization of uncertain inference methods from examples of desired behavior. This includes backpropagation and other optimizations of artificial neural networks.
- Machine learning and optimization of planning and search methods from examples from desired behavior. This includes reinforcement learning and optimization of recurrent neural networks.
- Machine learning and optimization of AI algorithms without examples of desired behavior ("unsupervised learning"). This includes clustering, principal-components analysis, latent semantic analysis, generative adversarial networks, and evolutionary algorithms.
- Reasoning methods that imitate those of humans or groups of humans. This includes implementations of a wide range of psychological theories.

Like other software, AI software can be assessed by several kinds of metrics:

- Accuracy of its logical reasoning. This is usually applied to classification tasks, and two classic metrics are correctness in what it classifies (precision), and completeness in what it classifies (recall).
- Accuracy of its numeric inferences: Average closeness to the correct answers using some error metric.
- Speed of its reasoning.
- Storage space required for its reasoning.
- Robustness in handling errors in its input
- Ability to explain its results and how it got them.
- Similarity of the results to human reasoning.

All these metrics bear on ethics. Most are predominantly quantitative. Thus they can be a basis for utilitarian ethics, or a basis for deontological ethics if we assign labels to ranges of numbers and refer to labels. However, the last two metrics above are more qualitative and need different assessment techniques.

Possible improvements to the ethics of AI algorithms

Ethical issues with AI software can be mitigated in several ways. The major ones are considered here.

Putting humans in the decision-making loop

A key worry with AI software is whether it can be trusted to think and act substantially as humans do, on the assumption that

humans are generally more ethical than machines since humans have higher-level goals. Subissues are whether AI systems can know everything that humans do to make decisions and whether they will reason similarly to humans with the same information. When these are concerns with military decisions, particularly those about lethal force, humans should be involved ("in the loop"); for instance, humans may know additional reasons that the AI does not as to why civilians are more likely to appear in a combat zone. Teams of carefully selected humans could also provide more diversity of points of view than AI could. AI could then serve an advisory role, recommending courses of action that could be overruled by human superiors. Many battle-management systems using AI are like this today.

However, such "hybrid" man-machine systems are not necessarily more ethical than machines alone (Cummings, 2021). Human personnel may not have access to the potentially huge amount of data and options that software might have, and so might make worse decisions than the software. Humans also have biases which can cause them to make bad decisions. These include well-known flaws in reasoning (Kahneman et al., 1982) such as a tendency to predict what they have seen before, something dangerous in military conflict where deception is often involved. Humans can also be influenced by propaganda, and can have deliberate unethical intentions. So putting humans in a decision loop will not necessarily ensure more ethical behavior.

Testing of AI systems

Some ethical problems with AI systems can be mitigated with proper testing. Software is complex and can easily contain harmful mistakes or flaws, that might cause lethal force to be used when the software designers did not intend so. Work on critical software has developed many testing methods to find bugs and flaws. Since most systems have too many possible inputs to test them all, sampling methods are essential though not guaranteed to find all bugs and flaws. A popular technique is "fuzzing" which tries small variants of tested input patterns to see if unusual effects occur.

Still, flaws in software are found all the time after it is released, and some of them can cause harm. Flaws are not always quickly reported publicly or quickly fixed after discovery (Lidestri and Rowe, 2022). Inadequate testing is common since incentives are weak for vendors to thoroughly debug before releasing software, and some vendors wait for users to find most bugs for them. It is difficult for users to recognize many bugs by themselves; many safety-related features in software are invoked rarely, so users cannot tell if they work properly. Nonetheless, many software vendors are conscientious, and voluntarily search for bugs.

Testing of AI machine-learning methods such as unsupervised learning that make random choices is particularly difficult. Such methods may give different answers when trained at different times on the same data, much less on different data. A solution is "cross-validation" where systems are trained to build models on random subsets of the data, and a consensus of the trained models taken as the result.

Explanation facilities: Inference

Lack of flaws alone is not enough to claim that AI software has acted ethically since its design may have other weaknesses. This is especially important with targeting, which can require careful judgement. Explanation capabilities can show how the software made its decisions, as a form of "transparency." Explanations also help debugging and provide legal justifications of AI (Atkinson et al., 2020). For software that does logical reasoning, an explanation can show the input data and the sequence of logical inferences made with it. For instance, if software identified a vehicle as hostile, an automatically generated explanation can show which features of the vehicle were relevant and what inferences supported the conclusion that it was hostile. Many AI systems that do logical reasoning make only a few logical inferences for a conclusion, and a trace of those will not overwhelm humans. Even better, we can allow users to ask "why" questions for particular conclusions made about the data that will identify just the data and inferences used. For instance, a system may conclude a vehicle is hostile if it has markings particular to an adversary and is in a location known to be controlled by an adversary.

For AI that does numerical calculations, explanation of decisions is harder. Typically such systems check whether the result of a calculation is over a threshold. The calculation is usually far too complex to explain to humans, especially with artificial neural networks. This raises problems for ethics because incomprehensibility prevents easy justification of the method. Some work on neural networks has tried to explain conclusions better; for instance, we can measure the impact of each factor or network feature on the complex mathematical function. However, this may not help much because often the correlations between factors matter more than the factors individually, and there are many possible correlations. To address this, some approaches try to identify larger parts of a neural network that have more impact on a conclusion, called areas of highest "salience" (Jacobson et al., 2018). However, this may not provide a good explanation either.

Explanations for military data could require revealing sensitive or classified data, such as data obtained by secret equipment. A less revelatory method may be to provide unclassified "precedents" for the case being explained. If they predominantly demonstrate the same conclusion as the case, the precedents and their reasoning can be presented. A challenge of this is defining similarity between cases: Some differences

should be given higher weights based on machine learning from examples.

Explanation facilities: Planning

AI can also be used to plan military operations. If unethical operations such as deliberate targeting of civilians are planned by AI, the result will be unethical regardless of the accuracy of the targeting software. Unfortunately, many planning systems are focused only on sensors, weapons, and logistics.

The ethics of plans generated by AI methods can be improved by calculating and displaying their ethical factors explicitly, such as possible civilian casualties of a plan or the risk of exerting disproportionate force. As with inference, explanations of plans can enable scrutiny and easier detection of ethical issues, up to some limits of complexity (Ananny and Crawford, 2016). Helpful explanations for logically-generated plans can reference preconditions, postconditions, and priorities on actions. "Why" questions about actions can be answered by relative costs and benefits, or by preconditions in a hierarchy of goals. But complex numerically-based plans can be hard to justify. Explaining targeting may require not only analysis of the costs and benefits of each target but the resources available and the logistics of getting them to the targets, and the tradeoffs can be complex. A simpler numeric model that can explain a similar plan can help, as for instance a Bayesian conditional-probability model rather than a deep neural network.

Looking for biases

AI systems can perpetuate unfair biases, particularly when they are developed using machine-learning methods on complex data. For instance, an AI system may be trained on U.S. data in which friendly forces were tall, and thus be more inclined to identify short people as combatants; or it may be trained on indicators of aggression seen in one part of the world, like maneuvers along a frontier, that may not occur elsewhere. Bias is particularly troublesome for AI systems because the bias may be deeply hidden in a large amount of data and no one may be aware of it. Some of these situations exemplify a well-known problem of statistical sampling of getting a representative sample of input. If important types are underrepresented in the raw data, data can be duplicated, or frequent types can be reduced in number (subsampled). Better transparency of systems by explanation tools can also help the analyses of their biases.

Automated ethical reasoning

Another way to improve the ethical behavior of AI software is to design the AI itself to use explicit ethical principles or criteria such as those of (Galliott, 2021). For instance, the principle of avoiding threats to civilians can be modeled by building a separate neural network that calculates the expected number of civilians to be harmed near a target based on intelligence data (Devitt, 2021) provides a start at a set of implementable principles. People seem to understand deontological ethics more easily than utilitarian ethics, so the principles will be easier to understand and justify if expressed as if-then rules. They will require setting thresholds on probabilities and other quantities, so designers must be prepared to argue why a 0.6 probability of killing a civilian is acceptable. Nonetheless, automated ethical principles could be better than human decision-making since they can avoid emotional responses to particular nationalities, ethnicities, political groups, or religions and thus could judge threats more objectively.

Recommendations

This article has discussed several ways to improve the ethics of AI systems, but the most important is transparency of their operations in the form of explanations of what they are doing. Thus, ethical AI methods should be simple to explain and easy to justify. Numerical AI methods like artificial neural networks are more likely to be problematic because the complexity of their calculations makes them difficult to explain. Methods requiring long logical reasoning chains of if-then rules cause similar problems. Numerical methods also often require thresholds for action (like the speed of a missile to entail a response) which can be difficult to justify, and this is especially a problem for decision-tree and support-vector methods. Unsupervised machine-learning methods are also problematic because they are hard to control.

These issues mean it is also important to reveal the algorithms and key details of AI software used for military applications so that potential ethical risks can be identified. Some ethical issues can also be monitored automatically from within AI software, such as by estimating the casualties of a course of action and using that in recommending decisions.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was supported in part by the U.S. National Artificial Intelligence Institute, the U.S. Department of Veteran's Health Affairs, and the U.S. Department of Defense Office of Research and Development.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views expressed are those of the author and do not represent the U.S. Government.

References

Ananny, M., and Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* 20, 973–989. doi: 10.1177/1461444816676645

Atkinson, K., Bench-Capon, T., and Bollegala, D. (2020). Explanation in AI and law: past, present, and future. *Artif. Intell.* 289, 103387. doi: 10.1016/j.artint.2020.103387

Cummings, M. (2021). "The human role in autonomous weapons design and deployment," in *Lethal Autonomous Weapons*. eds Galliott, J., MacIntosh, D., and Ohlin, J. (Oxford: Oxford University Press), 273–287.

Devitt, S. (2021). "Normative epistemology for lethal autonomous weapons systems," in *Lethal Autonomous Weapons*, eds Galliott, J., MacIntosh, D., and Ohlin, J. (Oxford: Oxford University Press), 237–257.

Emery, J. (2021). Algorithms, AI, and the ethics of war. *Peace Rev.* 33, 205–212. doi: 10.1080/10402659.2021.1998749

Galliott, J. (2021). "Toward a positive statement of ethical principles for military AI," in *Lethal Autonomous Weapons*, eds Galliott, J., MacIntosh, D., and Ohlin, J. (Oxford: Oxford University Press), 121–135.

Jacobson, V., Li, J., Tapia, K., and Morreale, P. (2018). Visualizing neural networks for pattern recognition. *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence.* (New York, NY: Association for Computing Machinery), 18–22.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.

Lidestri, M., and Rowe, N. (2022). Quantifying the milestones of cyber vulnerabilities. Proceedings of the 21st International Conference on Security and Management, Las Vegas, NV, US.

Parkin, S. (2015). *Killer Robots: The Soldiers that Never Sleep*. In: AI: the Ultimate Guide | Weapon. Available online at: https://www.bbc.com/future/article/20150715-killer-robots-the-soldiers-that-never-sleep (accessed June 20 2022).

Rowe, N. (2022). Algorithms for artificial intelligence. *IEEE Computer* 55:7, 87–102. doi: 10.1109/MC.2022.3169360

Tsamado, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., et al. (2021). "The ethics of algorithms: key problems and solutions", in L. *Floridi (ed.), Ethics, Policies, and Governance in Artificial Intelligence, Springer* 97–124. doi: 10.1007/978-3-030-81907-1_8

frontiers | Frontiers in Big Data

TYPE Brief Research Report
PUBLISHED 30 September 2022
DOI 10.3389/fdata.2022.992620

OPEN ACCESS

EDITED BY
George Lucas,
United States Naval Academy,
United States

REVIEWED BY
Hilde G. Corneliussen,
Vestlandsforsking, Norway
Katherine Chandler,
Georgetown University, United States

*CORRESPONDENCE
Kelly Fisher
kell.jh.fisher@gmail.com

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 12 July 2022
ACCEPTED 16 September 2022
PUBLISHED 30 September 2022

CITATION
Fisher K (2022) The role of gender in providing expert advice on cyber conflict and artificial intelligence for military personnel.
Front. Big Data 5:992620.
doi: 10.3389/fdata.2022.992620

The role of gender in providing expert advice on cyber conflict and artificial intelligence for military personnel

Kelly Fisher*

Department of Social Dynamics, Peace Research Institute Oslo, Oslo, Norway

This article draws upon original qualitative interview data with Norwegian male and female cyberengineer cadets at the Norwegian Cyber Defense Academy, who could in the future be working with AI-enabled systems in a variety of positions throughout the Norwegian military. The interviews explored how these cadets feel they as cyberengineers will be perceived in their future positions in the military, what challenges they feel they may face, and how gender may play a role in this. Different cyberengineers expressed concern about being able to communicate the cyber domain to their non-technology specialist colleagues due to the increasing complexity of new technologies. Gender appeared to be playing a role in this concern as the women interviewed expressed specific concerns that they feel as women, that they do not fit the stereotype of who is a cyberengineer, while some of the men felt that as cyberengineers they were seen as embodying a nerd masculinity, and that these gendered perceptions has implications for how they feel others perceive their competence levels. The findings from this article highlights gendered hierarchies in the military and the need for military institutions to focus on developing communication skills among those working with cyber operations. As the role of cyber is expected to grow in military operations, cyberengineers will need to find ways of communicating effectively with non-specialists—especially as complex AI-enabled systems are introduced. Finally, this paper argues the need for military institutions to take gender into account for this training and need for gender-sensitive policies.

KEYWORDS

AI, gender, cyber, military, communication, masculinity

Introduction

The rapid development of new technologies in society is resulting in many changes in how warfare is conducted (Feickert, 2021). One of those changes is the increasingly large role that cyber operations play, whether during actual warfare or in gray-zone conflict (Bilal, 2021). Due to the broad scope of what is cyber operations, definitional clarity is challenging and with no generally agreed upon definition of what cyber operations are (Dinstein and Dahl, 2020), or the cyber domain for that matter (Ringas et al., 2014). For the purpose of this paper, cyber operations are defined as an attack by an actor

(nation, non-state actors) upon another's cyber capabilities (Dinstein and Dahl, 2020), with the cyber domain being defined as including computers, networks, and anything else connected to the internet and communication capabilities (Ringas et al., 2014, p. 58). While still in its infancy, Artificial Intelligence (AI) is already playing a role in cyber operations, and this is expected to grow in the years ahead, both for offensive and defensive purposes (National Security Agency, 2021; Helkala et al., 2022).

A new challenge emerging from the increased complexity of these technologies is that troops and commanding officers and non-technology specialists in the military often have minimal understanding of the cyber domain (Jøsok et al., 2017). As a result, cyber operators have a greater responsibility to communicate effectively what is happening in the cyber domain to their commanding officers and fellow troops (Knox et al., 2018), especially as they will need to work closely during "multi-domain operations" (Feickert, 2021). A growing field of research has highlighted the importance of good communication skills for cyber professionals, and the ability to explain ongoings in cyber domain to their less tech-savvy colleagues (Dawson and Thomson, 2018). Furthermore, this poses questions of how the implementation of AI-enabled systems will contribute to the challenge of communicating complex technologies, as AI further obscures understanding how these technologies work (Ellis and Grzegorzewski, 2021).

Possible uses of AI in cyber operations may include programs that detect and then respond to malicious activity in the military's networks, and at a rate faster than humans could (Helkala et al., 2022). While there is enthusiasm for the use of AI, there are also concerns about unintended consequences resulting from its use (Ellis and Grzegorzewski, 2021). As many at the top of the command hierarchy would be held accountable for unintended consequences of AI use, there is likely to be hesitancy about deploying these AI-enabled systems (Helkala et al., 2022). Concerns about unintended consequences from AI exists across a number of different sectors, including both military and non-military (Steen et al., 2021). As these types of systems can offer important advantages in cyber operations, cyberengineers will need to be able to explain these systems in a way that ranking officers can understand.

However, it is not only an issue that cyber professionals need to be able explain these technologies to their colleagues, but also there is the matter of how these cyber professionals are viewed and perceived by their colleagues. Expertise and those seen as experts is relational (Collins and Evans, 2007), meaning that it is also a matter of how individuals are perceived regarding their level of expertise. Many factors can play a role in how someone is perceived, including gender (Ore, 2018). This raises a question of how gender might play a role in how cyber specialists feel they are perceived. Research has shown the gender biases that exist against female experts in a number of fields (Greve-Poulsen et al., 2021, p. 2), including in cybersecurity (Frieze and Quesenberry, 2019). While there is research outlining the importance of good communication skills amongst those working in cyber operations, including in the Norwegian military's Cyber Defense (Knox et al., 2018), few of these have taken into account how gender may play a role in this (Ask et al., 2021; pp. 33–35).

To address this knowledge gap, original qualitative interviews were carried out with cadets at the Norwegian Defense Cyber Academy. These cadets are in the final year of their education and will be deployed throughout the Norwegian military to support cyber capabilities. The cadets and types of tasks they will be working with are those in which AI-enabled programs may soon come to play a role, providing an opportunity to understand what challenges may exist for cyber cadets and how military training and educational institutions can try to address this issue. The main questions explored in this article are how these cadets feel they as cyberengineers will be perceived in their future positions in the military, what challenges they feel they may face as cyberengineers, and how gender may play a role in this. The implications of exploring perceptions of current students allows for institutions to explore how these perceptions align or differ from future working situations, and then aim to better prepare their students for future realities trainings and curriculum (Sipe et al., 2010; p. 345).

Norway's military presents an interesting case as its military is often praised for its efforts of having a gender balanced and inclusive military. Since 2015 Norway has had universal conscription for both men and women (Jakobsen, 2021) and in 2020 19% of Norwegian military personnel were women with 33% of all conscripts being women (Forsvaret, 2021). Despite this, different studies have been carried out showing the way in which women still face barriers to inclusion in the Norwegian military (Kvarving, 2019). However, little of this research has focused on female cyber cadets in the Norwegian military. Drawing from research findings on women working in the cybersecurity industry and IT field more broadly globally (Frieze and Quesenberry, 2019) and in Norway (Corneliussen, 2021), we can see that women face gendered stereotypes of who is seen as being technically competent. As Corneliussen (2021) found in research on women working in ICT, most of these women perceive and experience that technology is something seen as masculine, and an environment in which they face different barriers to inclusion. The findings from these interviews aim to contribute further knowledge both to gender dynamics in the military (Enloe, 1989) as well as those working with technology (Wajcman, 2000), and in the specific case of Norwegian cyberengineers, where those two fields overlap. Finally, this article highlights the relevance and importance of gender in understanding not only women's experiences, but men's experiences in the military (Christensen and Kyed, 2022), and aims to build upon a growing field of literature examining how new and emerging technologies are disrupting

and reenforcing gendered hierarchies in the military (Clark, 2018).

In the next section the methods carried out for this project are described. Following this the results based upon the qualitative interviews are presented. The final section is a discussion, and the paper concludes with recommendations for future research.

Methods

This paper is based upon semi-structured interviews with cadets at the Norwegian Defense Cyber Academy. Thirteen cadets were interviewed who are in the final year of their bachelor's degree in cyberengineering, which is about 1/3 of the class. The cyberengineering program is a combination bachelor degree where students receive training in telematics, cyberengineering, and military leadership. Ten of the cadets were men, and three were women. This represents a similar gender ratio of cadets studying cyberengineering at the Cyber Academy, where each class has about 50 students, and where usually between 15 and 30% of each cohort in recent years has been women.

Cyberengineering cadets generally are deployed across the whole military and may work in a number of roles, from maintaining radios and communications for field units, to working at the main office for the Norwegian Cyber Defence Force in Lillehammer. Cadets were chosen as AI-enabled systems use in the Norwegian military currently is limited or non-existent, and these cadets will likely be working with such systems or overseeing others using them in their future military career. Speaking with cadets rather than currently deployed cyberengineers presents opportunities to explore how they perceive their future roles, which can provide insights for training institutions on how they can better prepare their cadets for the realities of the field, which may differ from their perceptions (Sipe et al., 2010).

This study was approved by the Defense Force, and participation was completely voluntary. Cadets were sent initial information about the project and participation *via* email through their course instructor, and interested cadets then emailed back. All cadets received a consent form and were informed of their rights in line with the regulations carried out by the Norwegian Centre for Research Data (NSD).

Semi-structured interviews were carried out as they allowed me to maintain some order in the interview, while also being able to explore themes that emerged during the interview (Morris, 2015). Semi-structured interviews also enable a more conversational dynamic, where the interviewer asks questions but where the participants are able to express themselves as they desire (Morris, 2015, p. 3). As someone whose Norwegian competence is only moderate the interviews were carried out in English, which raises important questions about possible language-related challenges. My interview guide was designed with this under consideration, and while most cadets were comfortable speaking in English, the cadets were given the choice to speak in Norwegian if ever they were uncertain of how to express something.

As the scope of this project was focused on exploring themes rather than generalizability (Mcguirk and O'Neill, 2016), 13 interviewees provided enough data for a meaningful analysis and exploration of the topic. All of the interviews lasted at least an hour, with several lasting over 90 min, which provided over 15 h of interviews to transcribe and analyze. The interviews were then analyzed by using a thematic analysis, which allowed me to identify themes in the qualitative data and can be helpful when analyzing data focusing on participants' "experiences" and "understandings and perceptions" (Clarke and Braun, 2016; p. 88). As my project aimed to explore what challenges the cadets felt they may encounter when working in the military, and what influence gender may have, a thematic analysis was well suited for exploring the research question.

Research ethics were taken into consideration at every step of the project, including safely handling the data and anonymizing the participants, and also reflexivity from the researcher (Dowling, 2016). Reflexivity meant that I was paying attention to my own positionality, but also how I interacted and engaged with the data as I was analyzing it. This type of awareness also contributed to ensuring that the research was produced in a rigorous and trusthworthy manner.

Results

In this section I present interview excerpts to show the main themes that emerged from my thematic analysis. The themes presented include (1) Participants' perceptions that others in the military don't understand cyber; (2) Effectively communicating cyber to non-cyber; and (3) reflections on gendered perceptions of technologies and its impact on female cyberengineers. In the discussion section I relate these themes back to the broader fields of military ethics and gender studies. Pseudonyms are used here to anonymize the identity of participants.

Participants' perceptions that others in the military don't understand cyber

Several cadets expressed that at a broader level across the military there was a lack of understanding about the role that cyber capabilities play in the military.

> Morten: I think many people are not aware of how badly things can go, or how vulnerable systems are. So I don't think that cyberwarriors get enough credit, and often it is understandable, because when you are defending a network, it is not something that everyone physically sees, so it is difficult to understand everything we are doing. Fighter pilots by comparison, it is much easier to acknowledge, and it kind of has more prestige. If you shoot down an enemy aircraft, it is something that you see with the physical eyes, but something in cyber space can be really difficult to understand for normal people.

Others would also state that in addition to a possible lack of understanding, they felt as though some units in the military devalue the importance of cyber. When asked about stereotypes that might exist about cyberengineers in the military, Petter would share he felt cyberengineers were seen as the nerds with less prestige in the military, and how he thought this might impact how others see cyber.

> Petter: I will be deployed with field units that have no security professional, other than us cyberengineers, and there is a few of us in each battalion, and there I think we have this nerdy, overly anxious stereotype, that we are the guys who complain that everything they do is unsafe. Sort of a necessary evil. We are sort of the outsiders there. Everyone else is leadership, which is hard work, or you know the guys in infantry, like sharp shooters, and I understand that we can be annoying when we come up and tell them that they don't use their cell phones right.

While Petter shares here that one of the challenges is that other units don't take cybersecurity seriously, and how might stereotypes of cyberengineers as nerds maybe played a role in this, Julia shared that other members of the military are starting to take cyber more seriously.

> Julia: I think that people take it more seriously after the attack on Stortinget (Norwegian Parliament) and seeing what an attack can do. But I also think it is misunderstood, because cyber is so broad, and most people think of it as a computer and internet, but it is much more. Communications, satellites, radios, and much more.

Here Julia references a prominent hack that took place against the Norwegian Parliament in 2021 (Stolt-Nielsen and Lysberg, 2021), underscoring that cybersecurity and cyber capabilities can have a significant impact on Norwegian security. However, similar to other comments shown, Julia feels that there is a general misunderstanding of what is the cyber domain. This underscores the need for cyberengineers to be able to communicate what is ongoing in the cyber domain to other members of the military, a theme which many of the cyberengineers themselves pointed out, and which is the next theme I turn to.

Effectively communicating cyber to non-cyber experts

When asked what skills are needed for cyberengineers, there were a number of responses that emerged, including general technical competence, creativity, and the need to have good communication skills. As Lars shared:

> Lars: I think it is important to have a good understanding of ethics. As a cyberengineer you have more understanding of what the technology is, so it is important to be able to communicate to other people in a way so that they can understand.

As Lars highlights, as a cyberengineer not only do you need to be able to communicate in a way that people understand, but also as a cyberengineer you need to be able to explain the ethics associated with the technology (Ellis and Grzegorzewski, 2021) a point returned to in the discussion more fully.

Anne would speak about the importance of good communication skills. Yet when asked if she felt that cyber had prestige in the military, she would share:

> Anne: Absolutely not, or not yet at least, and that is something we have talked about quite a bit in our studies from the beginning. We have to dare to speak up, and are likely going to meet resistance, because we are going out as specialists, and not leaders. We have to advise them on something they know nothing about, so it is possibly easier to not consider what we are saying, and our job is trying to describe how what is happening in the cyber domain is important to everything else that is happening.

From an institutional point of view, these comments present important insights into what types of skills and training should be included for cyberengineers, and I return to this in the discussion after presenting the final theme from my analysis.

Reflections on gendered perceptions of technologies and its impact on female cyberengineers

Gender as a theme would come up first with the women I spoke with when asking the cadets about stereotypes about cyberengineers in the military.

> Anne: I think the stereotype is the typical nerd, with glasses and head buried deep in the computer. I think that is still what most people think, and when I tell people I am doing this, they are like, but oh you are a girl, so that is also something that hasn't changed.

Anne was among the first I interviewed, and when I spoke with another of the women in the program, Sara, she also said she didn't feel as though she fit the stereotype of who is a cyberengineer. When I asked if it had anything to do with being a woman she replied:

> Sara: Both that (being a woman), and also that I am not a gamer really. I feel like those who are gamers fit the stereotype better.

While both Anne and Sara expressed that they didn't think they fit the stereotype of who is a cyberengineer due to their gender, neither of them felt that being a women had an impact on them being treated differently in the military. However, Julia expressed that she thought there were moments where she was being treated differently because she was a woman.

> Julia: Many guys, they don't understand that women also know stuff about computers. And I have experienced it myself during the exercise when we had cyber operations, I had to be really patient, because, they expected less of me than the other guys.

Here we see that Julia feels that because of gendered stereotypes about technology (Corneliussen, 2021), she is seen as less competent when it comes to cyber operations. When speaking with some of the male cadets about gender equality in the military and amongst cyber operators, many of them spoke about the high level of gender equality in the Norwegian military and in cyberengineering. However a few of them did highlight that due to broader societal ideas about technology, this may lead to gendered stereotypes.

> Jens: Even me, I don't naturally assume that a woman would be interested in gaming on a PC, so that is a kind of stereotype, that isn't explicitly military, but that is in most of the society.

Interestingly, Jens expresses that it is gendered stereotypes about technological competence, and not about the military that in Norway might create barriers for female cyberengineers. However, as will now be shown in the discussion, gendered dynamics were also at play for the male cyberengineers.

Discussion

As Jøsok et al. (2017) highlight, cyber applications in the military "distort" military structures, as those lower in ranks often have higher technical competence than their officer (p. 497). My interview findings show that many of the cyberengineers feel challenged by this disruption of hierarchy, and they reflected on how they feel that in the broader military, few understand what the cyber domain is or take it seriously. This further adds to what the cadets feel is a challenge they will encounter when working in the military, which is explaining cyber-related challenges and topics to their commanding officers and fellow soldiers (Knox et al., 2018). This has important implications for military effectiveness, as the cyber domain will increasingly play a vital role, and there will be a need for good communication between "operator and commander... in order to communicate efficiently to support each other's sensemaking" (Jøsok et al., 2017 p. 493). The implementation of AI-enabled systems will add to this challenge of effective communication, as AI further obscures understanding how these technologies work (Ellis and Grzegorzewski, 2021). Yet as Lars comments shows about the ethics of these technologies, he not only feels a burden to communicate the cyber domain, but also to communicate the ethical challenges associated with It. As the use of AI is integrated into cyber operations in the future, it will be crucial that militaries focus on developing the skills to operate and understand AI systems, but also still focus on the development of good communication skills among cyber operators as the personnel will remain crucial despite advanced technologies (Ellis and Grzegorzewski, 2021).

Additionally this article has highlighted the role of gender in relation to effective communication of the cyber domain in the military. As Corneliussen (2021) found with women working in ICT in Norway, "negotiating their belonging" (p. 48) is often more difficult for women than men due to stereotypes about who is good with technology. In an organization such as the military, often seen as a gendered and a masculine institution (Kvarving, 2019), overlapping factors appear to play a role in how these women see themselves as fitting in. While Julia was the only woman who felt as though this had negative consequences, the perception among the women interviewed suggests that they are aware of the gendered institution they exist within. As the Norwegian military currently is made up of about 30% women, it is likely that those who these women would be interacting with, and communicating with about the cyber domain, would likely be men. Experiences of discrimination and perceptions of self which are tied to societal stereotypes can possibly contribute to uncertainty, creating extra barriers in the everyday tasks and assignments these women may work with. The challenge this provides to military institutions is to continue to focus on how they can try to create more gender-neutral perceptions of technology and technological

competence, and to take into account how these assumptions may impact female cyberengineers.

Gendered perceptions appear to have an impact on the men as well. None of the men interviewed expressed that they didn't fit the stereotype of who is a cyberengineer, further highlighting the gendered nature of technology and the military and of who feels they fit the stereotype. However, their comments about how they felt cyberengineers were viewed more broadly in the military illustrate their perceptions of masculine hierarchies in the military (Christensen and Kyed, 2022). As Petter shared, he felt that the cyberengineers were seen as overly anxious, reflecting that perhaps they are seen as embodying a "geek masculinity" (Salter, 2018) which is marginalized within the military. Studies on masculine culture in different military contexts have highlighted the way in which different units within the military can construct "hegemonic" ideals of masculinity for their unit, and also how they feel they may be marginalized or looked down upon by other units in the military (Clark, 2018). Based on the comments of Petter and other men I spoke with, they feel that within the broader military cyberengineers are seen as embodying a nerdy masculinity, which for them they feel creates challenges for how seriously they think they will be seen. New technologies being embraced by militaries globally have, and will continue to change the way in which warfare is conducted. From these men it would appear that the cyber domain sits in an arena of tension, one in which it might be looked down up by other units, but one that also will continue to play an increasingly vital role in multi-domain warfare. What is seen as masculine, and thus of value, is fluid, and has changed in the military before. A question that remains is if / when that might happen in the cyber domain in the context of the Norwegian military.

This article has presented initial findings on the types of challenges Norwegian cyberengineers feel they may encounter in the field, and how gender may play a role in this. It is important to acknowledge that these reflections are based upon their own perceptions, and limitations in the project design limit the extent to which this article can claim these women encounter actual biases. Further research is needed to explore the ways in which gendered assumptions and biases may be impacting the male and female cyberengineers during cyber operations. While the research on improving communication skills among those working with the cyber domain is growing, little of this research has taken into consideration gender. As training institutions seek to prepare these cadets for their future role, understanding how these cadets' perceptions and the role of gender in these perceptions, which may align with or differ from reality, can provide important insights for better training and education. These findings also have policy implications, and highlight the need for institutions and organizations to implement gender-sensitive policies that are attentive to local gender dynamics and set concrete goals and measurements for creating more inclusive environments in the military (Millar et al., 2021).

Data availability statement

The datasets presented in this article are not readily available because data is personal information about participants and is not to be shared. Requests to access the datasets should be directed to kell.jh.fisher@gmail.com.

Ethics statement

The studies involving human participants were reviewed and approved by Norwegian Centre for Research Data. The patients/participants provided their written informed consent to participate in this study.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

The funding for this project has come from the Research Council of Norway, under the SAMKUL program.

Acknowledgments

I would like to thank and acknowledge the Norwegian Cyber Defense Academy for its accommodation and allowing access to the students who were the participants for this paper. Additional thanks to Greg Reichberg and Kirsi Helkala who both provided support throughout the research and writing process.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ask, T. F., Lugo, R. G., Knox, B. J., and Sütterlin, S. (2021). "Human-Human communication in cyber threat situations: A systematic review," in *International Conference on Human-Computer Interaction*. Cham: Springer, 21–43.

Bilal, A. (2021). Hybrid Warfare—New Threats, Complexity, and 'Trust' as the Antidote. Available online at: https://www.nato.int/docu/review/articles/2021/11/30/hybrid-warfare-new-threats-complexity-and-trust-as-the-antidote/index.html (accessed on April 25, 2022).

Christensen, A-. D., and Kyed, M. (2022). From military to militarizing masculinities. *Int. J. Masculinity Stud.* 17, 1–4. doi: 10.1080/18902138.2022.2028428

Clark, L. C. (2018). Grim reapers: Ghostly narratives of masculinity and killing in drone warfare. *Int. Fem. J. Pol.* 20, 602–623. doi: 10.1080/14616742.2018.1503553

Clarke, V., and Braun, V. (2016). "Thematic analysis," in *Analyzing Qualitative Data in Psychology. 2nd edn*, eds L Evanthia and A Coyle. (Thousand Oaks, CA: SAGE Publications, Ltd), 84–103.

Collins, H. M., and Evans, R. (2007). *Rethinking |Expertise.* Chicago: University of Chicago Press.

Corneliussen, H. G. (2021). "Women empowering themselves to fit into ICT," in *Technology and Women's Empowerment*, ed E. Lechman. London: Routledge, 45–62.

Dawson, J., and Thomson, R. (2018). The future cybersecurity workforce: Going beyond technical skills for successful cyber performance. *Front. Psychol.* 9, 744. doi: 10.3389/fpsyg.2018.00744

Dinstein, Y., and Dahl, A. W. (2020). *Oslo Manual on Select Topics of the Law of Armed* Conflict: Rules and Commentary. Springer Cham. doi: 10.1007/978-3-030-39169-0

Dowling, R. (2016). "Power, subjectivity, and ethics in qualitative research," in Hay I. M. (ed.), *Qualitative research methods in human geography*. 4th edn. Oxford: Oxford University Press, pp. 29-44.

Ellis, D. C., and Grzegorzewski, M. (2021) *Big data for generals... and everyone else over 40*. MacDill Airforce Base, FL: The JSOU Press, 21–29.

Enloe, C. (1989). *Bananas, Beaches and Bases: Making Feminist Sense of International Politics*. Berkeley: University of California Press.

Feickert, A. (2021). *Defense Primer: Army Multi-Domain Operations (MDO)*. Washington, D.C.: Congressional Research Service.

Forsvaret (2021). *Armed Forces in Numbers*. Available online at: https://www.forsvaret.no/en/about-us/armed-forces-in-numbers (accessed on 15 June 2022).

Frieze, C., and Quesenberry, J. L. (2019). *Cracking the Digital Ceiling: Women in Computing Around the World*. Cambridge: Cambridge University Press.

Greve-Poulsen, K., Larsen, F. K., Pedersen, R. T., and Albæk, E. (2021). No gender bias in audience perceptions of male and female experts in the news: Equally competent and persuasive. *Int. J. Press/Pol.* 1–22. doi: 10.1177/19401612211025499

Helkala, K., Cook, J., Lucas, G., Pasquale, F., Reichberg, G. M., Syse, H., et al. (2022). "AI in cyberoperations: Ethical and legal considerations for endusers," in *Articial Intellgience and cybersecurity: Theory and Applications*. Thousand Oaks, CA: Springer.

Jakobsen, S. E. (2021). *Researchers Conducted a Gender Equality Experiment on 500 Recruits: Do Men Become More Open to Gender Equality by Sharing Dormitories and Tasks With Women in the Military*? Available online at: https://sciencenorway.no/gender-and-society/researchers-conducted-a-gender-equality-experiment-on-500-recruits/1869059 (accessed June 9, 2022).

Jøsok, Ø., Knox, B. J., Helkala, K., Wilson, K., Sütterlin, S., Lugo, R. G., et al. (2017). "Macrocognition applied to the hybrid space: team environment, functions and processes in cyber operations," in *International Conference on Augmented Cognition,* 486–500.

Knox, B. J., Jøsok, Ø., Helkala, K., Khooshabeh, P., Ødegaard, T., Lugo, R. G., et al. (2018). Socio-technical communication: The hybrid space and the OLB model for science-based cyber education. *Military Psychol.* 30, 350–359. doi: 10.1080/08995605.2018.1478546

Kvarving, L. P. (2019). Gender perspectives in the armed forces and military operations: An uphill battle. *Cultural, structural and functional factors that prevent or promote implementation of UNSCR 1325 in the Norwegian Armed Forces and NATO. PhD thesis*. Oslo: University of Oslo.

Mcguirk, P. M., and O'Neill, P. (2016). "Using questionnaires in qualitative human geography," in *Qualitative Research Methods in Human Geography. 4th edn*, ed I. M. Hay (Oxford: Oxford University Press), 246–272.

Millar, K., Shires, J., and Tropina, T. (2021). *Gender Approaches to Cyber Security: Design, Defence, and Response*. Geneva: United Nations Institute for Disaramament Research.

Morris, A. (2015). *A Practical Introduction to In-Depth Interviewing*. Thousand Oaks, CA: SAGE Publications, Ltd.

National Security Agency (2021). Artificial intelligence: Next frontier is cybersecurity. Available online at: http://www.nsa.gov/Press-Room/News-Highlights/Article/Article/2702241/artificial-intelligence-next-frontier-is-cybersecurity/ (accessed on June 15, 2022).

Ore, T. (2018). *The Social Construction of Difference and Inequality. 7th edn.* Oxford: Oxford University Press.

Ringas, E. T., Kertunnen, M., and Spirito, C. (2014). *Cyber Security as a Field of Military Education and study*. Joint Force Quarterly. p. 75. Available online at: https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-75/jfq-75_57-60_Tikk-Ringas-et-al.pdf

Salter, M. (2018). From geek masculinity to Gamergate: The technological rationality of online abuse. *Crime Media Cult.* 14, 247–264. doi: 10.1177/1741659017690893

Sipe, S., and Johnson, C. D., and Fisher. D. (2010). University students' perceptions of gender discrimination in the workplace: Reality versus fiction. *J. Educ. Bus.* 84, 339–349. doi: 10.3200/JOEB.84.6.339-349

Steen, M., Timan, T., and van de Poel, I. (2021). Responsible innovation, anticipation and responsiveness: case studies of algorithms in decision support in justice and security, and an exploration of potential, unintended, undesirable, higher-order effects. *AI and Ethics* 1, 501–515. doi: 10.1007/s43681-021-00063-2

Stolt-Nielsen, H., and Lysberg, M. (2021). *To dataangrep på tre uker på Stortinget: Kontaktnettverk, norske standpunkter og indre konflikter har etterretningsverdi, sier E-tjenesten*. Available online at: https://www.aftenposten.no/norge/i/G304k9/to-dataangrep-paa-tre-uker-paa-stortinget-kontaktnettverk-norske-standpunkter-og-indre-konflikter-har-etterretningsverdi-sier-e-tjenesten (accessed June 9, 2022).

Wajcman, J. (2000). Reflections on gender and technology: In what state is the art? *Soc Stud Sci.* 30, 447–464. doi: 10.1177/030631200030003005

frontiers | Frontiers in Big Data

TYPE Brief Research Report
PUBLISHED 20 October 2022
DOI 10.3389/fdata.2022.991459

OPEN ACCESS

EDITED BY
Kirsi Helkala,
Norwegian Defence University College, Norway

REVIEWED BY
Martti Lehto,
University of Jyväskylä, Finland
James Cook,
United States Air Force Academy, United States

*CORRESPONDENCE
Richard Dean
rdean@calstatela.edu

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 11 July 2022
ACCEPTED 20 September 2022
PUBLISHED 20 October 2022

CITATION
Dean R (2022) Lethal autonomous weapons systems, revulsion, and respect. *Front. Big Data* 5:991459. doi: 10.3389/fdata.2022.991459

Lethal autonomous weapons systems, revulsion, and respect

Richard Dean[1,2]*

[1]United States Naval Academy, James B. Stockdale Center for Ethical Leadership, Annapolis, MD, United States, [2]Department of Philosophy, California State University, Los Angeles, Los Angeles, CA, United States

The potential for the use of artificial intelligence in developing lethal autonomous weapons systems (LAWS) has received a good deal of attention from ethicists. Lines of argument in favor of and against developing and deploying LAWS have already become hardened. In this paper, I examine one strategy for skirting these familiar positions, namely to base an anti-LAWS argument not on claims that LAWS inevitably fail to respect human dignity, but on a different kind of respect, namely respect for public opinion and conventional attitudes (which Robert Sparrow claims are strongly anti-LAWS). My conclusion is that this sort of respect for conventional attitudes does provide some reason for actions and policies, but that it is actually a fairly weak form of respect, that is often override by more direct concerns about respect for humanity or dignity. By doing this, I explain the intuitive force of the claim that one should not disregard public attitudes, but also justify assigning a relatively weak role when other kinds of respect are involved.

KEYWORDS

lethal autonomous weapons systems (LAWS), artificial intelligence, military ethics, respect, conventional attitudes, revulsion

Introduction

Robert Sparrow has argued that since there is widespread moral revulsion at the idea of developing and deploying Lethal Autonomous Weapons Systems (LAWS), to do so would show disrespect for humans who have this feeling or attitude of revulsion (Sparrow, 2016). Given some (controversial) empirical assumptions, I think he may be right that respecting this conventional attitude provides a *prima facie* reason to reject LAWS. But keeping in mind a distinction between this kind of symbolic respect and a deeper and more foundational form of respect for persons, it can be seen that Sparrow's proposed respect for conventional attitudes only provides a fairly weak, *prima facie* reason for action. This is important, because it acknowledges the strong aversion that some people have to the possibility of developing LAWs, but simultaneously suggests limits on the argumentative role of this feeling.

Hypothesis

It is possible to accommodate the important moral intuition that one ought to show a kind of respect for widespread attitudes and conventions, without supposing that such attitudes provide an overriding reason to reject all use of LAWS.

Method

I will examine existing views on the role of respect for conventional attitudes on the moral permissibility of developing and employing LAWS, extract the key elements of these views, and argue that these views can be accommodated without a comprehensive ban on LAWS.

Method: Moral debate about LAWS

The longstanding trend toward using military weapons from great distances has accelerated in recent decades, with the use of artillery and aerial bombardment being supplemented by precision guided missiles and remotely piloted aircraft. In addition to the increasing distance between combatants and their targets, some weapons systems have the capability to operate in ways that are significantly autonomous, or independent of direct human control, the movement toward increasingly remote operation of lethal weapons and the movement toward systems capable of operating without direct human control have not fully intersected. Weapons capable of operating fully autonomously have either been defensive, like the US Aegis missile defense system (when on "auto special" setting), or they have been directed at targets other than humans, like Israel's Harop system, which is designed to destroy radar equipment. However, this historical separation of the autonomy of weapons systems and their lethal application to human targets apparently has collapsed recently, as the first uses of autonomous drones against human targets, without human oversight, reportedly have already occurred (Cramer, 2021).[1]

The anticipatory moral debate about developing and deploying such LAWS has been lively and passionate. It is widely agreed that the technological limits of current AI make it ethically unjustified to assign decisions about targeting humans to weapons systems, independent of human oversight—that is, to employ weapons systems against human targets with human controllers "out of the loop," to use current terminology. While AI may well distinguish between allied and enemy forces in many battlefield conditions, it is not yet as reliable as humans in making more subtle judgments about whether enemy combatants are *hors de combat*, or at identifying irregular forces, making judgments about whether civilians are actively supporting military operations, or foretelling collateral damage and civilian casualties. But it seems inevitable that the AI employed in LAWS will eventually become at least as good as humans at tasks like this. The question of whether there is a principled reason to prohibit the use of LAWS at that point is a matter of heated debate, and a number of lines of moral argument for and against their use have become firmly entrenched.

The most fundamental principled objection to the development and use of LAWS is that removing human control from the process of targeting and killing human beings would in some way show disrespect for humankind, or would be a failure to recognize and acknowledge human dignity (Asaro, 2012). The objection turns on a claim that respect for human dignity requires some kind of active recognition of the humanity of the target, a recognition of which machines are inherently incapable.[2] But a response to this objection also has become standard, namely that by this standard, LAWS should be no less acceptable than many weapons that have been in use for many years or even decades, such as cruise missiles or standard artillery (Jenkins and Purvis, 2016).

Proponents of the use of LAWS have developed their own influential argument, emphasizing that when LAWS become better than human operators at distinguishing between legitimate military targets and civilians, the lives saved will constitute such a positive consequence that the use of LAWS will not be morally wrong, but may instead be morally required (Arkin, 2010). This pro-LAWS argument also can be presented as a matter of respecting the dignity of the humans whose lives are saved, directly countering the argument that use of LAWS must be eschewed for the sake of respecting human dignity (Jenkins and Purvis, 2016).

The basic positions have been staked out for several years, with a good deal of the moral thinking on the topic of LAWS consisting more or less of modifications and reinforcements within this framework (Skerker et al., 2020; Bohrer, 2022; Kahn, 2022). In this kind of hardened rhetorical landscape, it is useful to try new approaches, and that seems to be the motivation for an argument offered by Robert Sparrow, in which respect and disrespect play a different role than in the standard anti-LAWS position.

Results: Sparrow's "conventional respect" argument against LAWS

Sparrow relies on an idea that what counts as respectful or disrespectful can depend on "social understandings" so

1 Although even simple weapons, such as land mines or IEDs, can operate without direct human control, I am following the parameters of the debate started by Sparrow and others, in focusing on lethal weapons systems directed at human targets that involve some discrimination or targeting with humans out of the loop. (Some "automatic" weapons like the Russian POM-3 mines may blur the line between automatic and more fully autonomous weapons, in that they do discriminate between human and other moving targets).

2 As an anonymous reviewer has noted, the response below presupposes that one can view a machine as distinct from the intention of its creator. So, besides the standard response I describe, a more radical response is possible, in which machines as artifacts can express a variety of attitudes of their human creators.

there is a "conventional element to our understanding of the requirements of respect" (Sparrow, 2016, p. 109). Sparrow's approach may circumvent the need to settle some of the standard, highly controversial questions about human dignity, respect, and their role in debates about LAWS. This is because instead of attempting to establish theoretically that the nature of human dignity and respect for persons requires a direct and personal engagement with any person who is affected by life and death decisions, it substitutes a different kind of respect, namely respect for conventional attitudes. To flout these widespread attitudes sends a message of disrespect to people who have them, by implying that their feelings or attitudes are unimportant.

Sparrow's strategy here places great weight on a supposed "widespread public revulsion at the idea of autonomous weapons" (Sparrow, 2016, p. 109), and that "Most people already feel strongly that sending a robot to kill would express a profound disrespect of the value of an individual human life" (Sparrow, 2016, p. 109). Sparrow admits that "it is possible that public revulsion at sending robots to kill people will be eroded as AWS come into use and become a familiar feature of war" (Sparrow, 2016, p. 116), but he regards this as a significant change from what he takes to be the currently prevalent attitude, that the use of LAWS would be an appalling example of failure to respect human dignity. Sparrow says, "the strength and popular currency of the intuition that the use of [LAWS] would profoundly disrespect the humanity of those they are tasked to kill is sufficient justification to try to establish such a prohibition" (Sparrow, 2016, p. 111).

Although I will grant, for the sake of argument, Sparrow's premise that there is widespread public revulsion at the thought of the use of LAWS, this is only a hypothetical concession, and in fact the premise is not strongly supported. Sparrow's main evidence is a 2013 survey, in which 39% of respondents said they "strongly oppose" the use of "robotic weapons that can independently make targeting and firing decisions without humans in the loop," and 16% said they "somewhat oppose" it (Carpenter, 2013). However, the survey did not ask why respondents held their views, so it is hasty to conclude that even the 55% of respondents who opposed the use of LAWS did so because of a feeling of repugnance. There are many other reasons why someone might oppose the use of LAWS against humans, including concerns about the technical adequacy of AI in targeting, a general pacifism, or a resistance to increasing the gap between nations with advanced military technology and those without. A 2021 survey by Human Rights Watch is more suggestive of some public unease or perhaps revulsion at the use of LAWS, though it is still inconclusive. Asked whether they support or oppose the use of LAWS (with the concept of LAWS being explained within the survey question), 41.9% of all respondents said they "strongly oppose" LAWS and 19.4% "somewhat oppose" it. This survey also asked respondents who opposed the use of LAWs for the reasons for their opposition. It found that 66.2% of all respondents who oppose the use of LAWS said, "They'd cross a moral line because machines should not be allowed to kill" (Human Rights Watch, 2021). There is room for concern about circularity—use of LAWS is wrong because it crosses a line into wrongness—but the idea of a "moral line" also is at least compatible with a feeling of moral revulsion. So, charitably, it may be that about 40% of all respondents (two-thirds of the 63% of respondents who opposed the use of LAWS) feel at least some unease or perhaps even revulsion at the thought of military use of LAWS. This is some potential evidence, but weak evidence, for the claim that there is a widespread feeling of revulsion toward LAWS.

Nevertheless, it is worth granting this contestable claim, to see how strong an argument against the use of LAWS can be formulated, using it as a premise. Not only does the 2021 survey mentioned above hint at some revulsion, but it also is undoubtedly the case that some commentators on LAWS display strong and deeply held feelings about the repugnance of allowing machines to make decisions about lethally targeting humans, and about the incompatibility of this practice with human dignity. Besides the authors Sparrow mentions (e.g., Gubrud, 2014), many other examples could be cited (e.g., Heyns, 2017). It is unclear what amount of public revulsion is needed in order to count as supporting a moral requirement of respecting the attitude. The answer to this question is neither obvious nor empirically resolvable, but would instead itself be a matter for moral argument. Instead of embarking on that project, I will grant hypothetically that there is a "strong" feeling of revulsion at the thought of using LAWS, and see what argument follows.

Sparrow does not simply take a feeling of moral revulsion to provide direct proof that some practice, like the use of LAWS, is wrong. That strategy, which is sometimes (probably uncharitably) attributed to Kass (1997), seems implausible.

Instead, Sparrow's argument works through an idea of societal conventions about what counts as respectful or disrespectful treatment. Although Sparrow does not go into great detail about the connection between feelings of revulsion and conventions regarding respect, his thought appears to be that feelings of revulsion about some practice are one indication that deeply held conventions or attitudes regarding respect are being violated. This is consistent with his example of a deep feeling of disgust in reaction to the mutilation of corpses, even though what counts as mutilation "is conventional and may change over time" (Sparrow, 2016, p. 109). Even though a specific way of treating corpses (cutting fingers off, eating parts of them, burning them) may be regarded as disrespectful according to the conventional standards of one society or a set of societies, the conventions of some other society might deem the same treatment respectful. But these ways of treating corpses genuinely are disrespectful in virtue of violating conventions, despite the conventions being mutable, because violating conventions about respect is a way to exhibit disrespect. If some society has a convention against touching strangers with one's left hand, then deliberately touching a stranger with

one's left hand in this society is disrespectful, whether the origins of the convention have to do with health and hygiene, religion, combat practices, or just superstition, and regardless of whether some other society holds any such convention. In the same way, Sparrow maintains that attitudes of revulsion at the thought of employing LAWS show that there is a widely shared conventional attitude that assigning decisions about taking a life to machines is disrespectful to human dignity, and that violating these conventions is a way of showing disrespect. "That the boundaries of such respect are sometimes—as in this case—determined by convention (in the sense of shared social understandings rather than formal rules) does not detract from the fact that it is fundamental to the ethics of war" (Sparrow, 2016, p. 110).

So, a simple representation of Sparrow's argument would be:

(a) If some type of action violates conventions of respect and disrespect, then this is a moral reason not to perform this type of action.
(b) The development and use of LAWS violates widespread conventions regarding respect and disrespect for humans, which require personal recognition and acknowledgment of the life being taken (and revulsion at the idea of using LAWS is evidence of this violation of conventions).
(c) Therefore, we have a moral reason not to develop and deploys LAWS.

This representation of Sparrow's argument is deliberately vague (in that it leaves open how compelling a reason is provided by respect for conventions), and examination reveals that if more properly specified, it is a sound argument, but that it also is limited in the ramifications of its (fairly weak) conclusion.

Discussion: Sparrow's argument and two levels of respect

There is something intuitively compelling about Sparrow's argument. It does seem morally problematic to flout conventions regarding respectful and disrespectful behavior, or to dismiss feelings of revulsion at possibly grave violations of some of these conventions. But the question is how much weight to give to these norms and attitudes, compared to other considerations involved in the possible use of LAWS.

In a paper directly responding to Sparrow's position, Purves and Jenkins acknowledge that "public aversion to a technology counts against its adoption," but they are quick to dismiss some of these attitudes, because the attitudes are not based on sound moral reasons (Purves and Jenkins, 2016, p. 396). They say that "public opinion can be swayed by an array of factors, only some of which are indicative of the moral truth of a matter," and that "ethicists should not be satisfied to let public opinion carry the day, especially in the absence of a robust moral distinction…" (Jenkins and Purvis, 2016, p. 396). In effect, they question premise 1 of the reconstruction of Sparrow's argument offered above—they claim that conventions of respect should only be accommodated if they are based on sound moral reasoning. Some thought experiments can be generated in support of this position. For example, suppose that some nation at war pleads with enemy forces to only use combatants of northern European descent in military engagements with them. "Please do not allow people of color to take our lives," they say, "This is deeply disrespectful of our traditions, which maintain that only people of our own race are worthy opponents." Intuitively, it seems that such a plea should carry no weight, because it is based on misguided, racist, moral ideals.

But, despite cases like this, it is hasty to conclude that we should respect only feelings and conventions that we think are based on sound moral reasons. Suppose that instead of pleading that no person of color should be deployed as a combatant against them, the nation at war pleads that ammunition used against them should contain no copper. "Our spiritual and religious convictions tell us that copper is impure, and contaminates our souls," they say. If ammunition were available that did not contain copper, and if its use were as effective as ammunition containing copper, then it seems that their revulsion at the idea of being killed by copper should carry moral weight, providing at least some moral reason to use non-copper ammunition. And this would be so, even if their aversion to copper seemed to be based on mere superstition, instead of any sound moral reasoning. For that matter, the example of avoiding touching strangers with one's left hand seems to be a real way of showing respect, even if the origins of the convention are disconnected from any current negative or positive effects.

Instead of saying that conventions of respect should carry weight only if they can be seen to rest on sound moral reasoning, a more nuanced account is needed. Our own choices, especially if they are deliberate policy decisions, not only lead to actions, but also send a message. In this sense, Sparrow is correct that "ethics is a realm of meanings" (Sparrow, 2016, p. 101). I can deliberately flout conventions of respect and send a message that those conventions are morally objectionable, or I can simply fail to consider them, and send a message that I disrespect those who hold the conventional attitudes. Or I can adhere to the conventions of respect even if I do not see their point, sending a message of respect for those who do regard the conventions as important. This picture, encompassing the communicative element of actions, allows for a subtler, more nuanced picture of the role of conventions. However, it is of little pragmatic help in telling us how much weight to give to some particular, controversial conventions or attitudes, such as the attitude that the use of LAWS is contrary to human dignity.

But a point about respect drawn from normative moral theory is useful here, and suggests that in the case of LAWS, respect for conventions or attitudes provides a relatively weak reason for action, which may well be outweighed by

other considerations. This can be seen by disambiguating the concept of "respect," which is used in many different ways in moral discourse.

Sparrow's argument centers on a specific type of respectful action, namely actions that have to do with adherence to or rejection of conventions. Such conventions can encompass matters like not touching strangers with one's left hand, or more profound matters, such as how corpses are handled or how lethal combat decisions are made. We can show respect by adhering to conventions, respect both for individuals affected by our actions and respect for those who hold the conventions. But there must be a rationale for recognizing and adhering to any class of duties, including these duties of conventional respect.

And this deeper rationale for recognizing various classes of duties is sometimes also described by using the word "respect," in a different sense. Kantian theories often make respect for persons, or for the dignity of humanity, the deep basis of many or all duties. So Wood takes respect for the dignity of rational nature as the basis of all duties to oneself or others (Wood, 2008). Darwall, whose views have a looser affinity with Kant's, identifies a type of respect he calls "recognition respect," and takes it that recognition respect for persons is the basis of our duties regarding what one person can morally demand of another (Darwall, 1977). Hill takes from Kant a moral requirement to treat all persons with "basic respect" as potential moral decision-makers, with this basic respect leading to more specific moral requirements (Hill, 2000). The strategy of grounding moral claims on a foundational respect for persons or for dignity is not restricted to moral theory, it is also displayed in important public statements, such as the United Nations' Universal Declaration of Human Rights, which grounds its specific human rights on "recognition of the inherent dignity" of all humans, and requires universal and equal "respect" for these rights (United Nations, 1948).

This foundational role that respect for persons or for human dignity often plays imbues the word "respect" with an aura of moral significance, even inviolability. After all, if all duties are based on respect for persons, then to fail to give this kind of respect is by definition wrong. Even if only some substantial subset of duties are based on respect for persons, then it would take powerful countervailing reasons to override these duties.

But a derivative set of duties are more likely to be defeasible, even if the duties go by the name of duties of "respect." Symbolic respect for persons, instantiated in actions such as trying to abide by conventions of respect, may be based on a foundation of deep respect for persons, but may be frequently overridden by other duties that also are based on foundational respect for persons. In fact, it appears that duties such as preventing loss of life, distributing goods and outcomes fairly, and maybe even being truthful, often outweigh symbolic respect for conventions, in a moral system based on a foundational respect for persons. There is *prima facie* reason to abide by a convention against touching strangers with one's left hand, but this reason becomes inert if the only way to save a stranger from a burning car is to pull her out with both hands. There is some moral reason to eschew ammunition containing copper in order to show symbolic respect for a society's conventional attitudes, but if the war can be ended more quickly and with much less suffering by using ammunition containing copper, the prohibition on copper falls away easily. And, in a military context, even the convention of adhering strictly to rank hierarchies, which is quite strong in modern militaries, can be outweighed when one is given orders that violate the standard rules of *jus in bello*, which can plausibly be seen as being based on respect for humanity.

While I do not claim that requirements of respect for conventions are so weak as to always be outweighed by any conflicting moral considerations, it does seem that they are often relatively weak *prima facie* duties, if one keeps in mind the distinction between duties of respect for specific conventions and the deep foundational respect that many ethicists take to be central to morality. If so, then Sparrow has identified a source of the intuition that conventional attitudes of repugnance toward the use of LAWS provide some reason to refrain from developing and using LAWS. But it is a weak reason, that does not obviously seem to outweigh other morally important factors, such as minimizing loss of life or self-defense against unjust attacks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This study was supported by PRIO, Peace Research Institute Oslo.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arkin, R. (2010). The case for ethical autonomy in unmanned systems. *J. Military Ethics* 9, 332–341. doi: 10.1080/15027570.2010.536402

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *Int. Rev. Red Cross* 94, 687–709. doi: 10.1017/S1816383112000768

Bohrer, A. (2022). *The Moral Case for the Development of Autonomous Weapons Systems*. Available online at: https://blog.apaonline.org/2022/02/28/the-moral-case-for-the-development-of-autonomous-weapon-systems/ (accessed June 20, 2022).

Carpenter, C. (2013). *US Public Opinion on Autonomous Weapons*. Available online at: http://www.duckofminerva.com/wp-content/uploads/2013/06/UMass-Survey_Public-Opinion-on-Autonomous-Weapons.pdf (accessed June 20, 2022).

Cramer, M. (2021, June 3). AI drone may have acted on its own in attacking fighters, U.N. says. *New York Times*.

Darwall, S. (1977). Two kinds of respect. *Ethics* 88, 36–49. doi: 10.1086/292054

Gubrud, M. (2014). Stopping killer robots. *Bull. Atomic Sci.* 70, 32–42. doi: 10.1177/0096340213516745

Heyns, C. (2017). Autonomous weapons in armed conflict and the right to a dignified life: an African perspective. *South Afr. J. Hum. Rights* 33, 46–71. doi: 10.1080/02587203.2017.1303903

Hill, T. (2000). *Respect, Pluralism, and Justice: Kantian Perspectives*, Oxford: Oxford University Press.

Human Rights Watch (2021). *Autonomous Weapons*. Available online at: https://www.ipsos.com/sites/default/files/ct/news/documents/2021-01/ipsos_global_advisor_-_lethal_autonomous_weapons_survey_-_nov_2020-jan_2021.pdf (accessed June 20, 2022).

Jenkins, R., and Purvis, D. (2016). Robots and respect: a response to Robert Sparrow. *Ethics Int. Affairs* 30, 391–400. doi: 10.1017/S0892679416000277

Kahn, L. (2022). Lethal autonomous weapons systems, jus in bello, and respect for human dignity. Cybersecurity and Privacy, a section of the journal Frontiers in Big Data.

Kass, L. (1997). The wisdom of repugnance. *New Republic* 32, 17–26.

Skerker, M., Purves, D., and Jenkins, R. (2020). Autonomous weapons systems and the moral equality of combatants. *Ethics Inform. Technol.* 22, 197–209. doi: 10.1007/s10676-020-09528-0

Sparrow, R. (2016). Robots and respect: assessing the case against autonomous weapons systems. *Ethics Int. Affairs*. 30, 93–116. doi: 10.1017/S0892679415000647

United Nations (1948). *Universal Declaration of Human Rights*. Available online at: https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed June 20, 2022).

Wood, A. (2008). *Kantian Ethics*. Cambridge: Cambridge University Press.

frontiers | Frontiers in Big Data

TYPE Mini Review
PUBLISHED 25 October 2022
DOI 10.3389/fdata.2022.991392

OPEN ACCESS

EDITED BY
Kirsi Helkala,
Norwegian Defence University College, Norway

REVIEWED BY
Grethe Østby,
Norwegian University of Science and Technology, Norway
Gregory Reichberg,
Peace Research Institute Oslo (PRIO), Norway

*CORRESPONDENCE
Mark Metcalf
mmetcalf@virginia.edu

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 11 July 2022
ACCEPTED 16 September 2022
PUBLISHED 25 October 2022

CITATION
Metcalf M (2022) The PRC considers military AI ethics: Can autonomy be trusted?. *Front. Big Data* 5:991392. doi: 10.3389/fdata.2022.991392

The PRC considers military AI ethics: Can autonomy be trusted?

Mark Metcalf*

McIntire School of Commerce, University of Virginia, Charlottesville, VA, United States

China's People's Liberation Army (PLA) is currently wrestling with the benefits and challenges of using artificial intelligence (AI) to enhance their capabilities. Like many other militaries, a key factor in their analysis is identifying and dealing with the ethical implications of employing AI-enabled systems. Unlike other militaries, however, as the PLA is directly controlled by the Chinese Communist Party (CCP—"the Party"), such considerations are conspicuously influenced by a definition of military ethics that is fundamentally political. This Mini-Review briefly discusses key tenets of PLA military ethics and then investigates how the challenges of military AI ethics are being addressed in publicly-available government and PLA publications. Analysis indicates that, while the PLA is considering AI ethical challenges that are common to all militaries (e.g., accountability), their overriding challenge that they face is "squaring the circle" of benefitting from autonomous AI capabilities while providing the CCP with the absolute control of the PLA that it demands—which is, from the CCP's perspective, a military ethics consideration. All Chinese translations are my own.

KEYWORDS

artificial intelligence, intelligentization, lethal autonomous weapons systems, military ethics, People's Liberation Army

Introduction

This paper discusses People's Republic of China (PRC) writings on military artificial intelligence (AI) ethics as they apply to the People's Liberation Army (PLA).

While an abundance of academic and military research about military AI ethics has been published in the PRC, there is currently not a publicly-available *official* PLA policy on the topic.

The PLA is unwilling to publish any materials that may provide potential adversaries insights into their specific considerations and plans for the use of new technologies (such as AI), as such information is considered state secrets. Instead, such writings exhibit a high degree of indirectness. For example, a PLA researcher, instead of unequivocally stating that the PLA should be concerned about the ethical issue of accountability when using lethal autonomous weapons systems (LAWS), will summarize research by *Western* scholars that express concern about the issue. This allows the author to avoid making a policy recommendation—which is the Chinese Communist Party's prerogative—while indirectly highlighting a concern about AI accountability. When evaluating PRC writings on military AI ethics, then, readers must frequently "read between the lines" (relying on multiple sources) to determine what is actually being proposed. They must not only consider

what is written, but also how and why specific issues are—and aren't—addressed [(Jullien, 1995), p. 93–115].

As a result, there are currently only a handful of PLA sources that, somewhat authoritatively, express PLA perspectives on military AI ethics and they form the basis of this paper.

An overview of PLA military ethics

To comprehend the PRC's views regarding military AI ethics, it is important to have a basic understanding how the PLA views the scope and role of military ethics; a perspective that is strongly influenced by the Party (Metcalf, Forthcoming)[1].

A consistent theme in PLA ethics writings is the importance of developing a "military ethics culture [that] guides soldiers' ethical self-awareness and moral self-discipline" [(Tang, 2016), p. 2]. The PLA considers military ethics to be a political matter that is guided by the Party; an issue that is rooted in Marxist ethics and CCP General Secretary Xi Jinping thought, that supports the development of "socialism with Chinese characteristics", and contributes to the Party's goal of a creating a strong military as a key element of a rejuvenated China [(Liu and Li, 2020), p. 74]. As a result, PLA writings about military ethics emphasize both the political goals and military benefits of military ethics and rarely engage in discussions of ethics for ethics' sake. In 2017, for example, the Party directed that the PLA should "Follow the Party! Fight to win! Forge exemplary conduct!"—an omnipresent saying in venues ranging from military newspapers to propaganda posters to military facilities. Elsewhere, a PLA political officer explains the strategic significance of military ethics using distinctly political terminology.

> The development of military ethics thus embodies the unity of scientific and revolutionary, the unity of theory and practicality. It is not only an important part of the revolutionary change of military culture, but also the historical memory of the nation, a concentrated embodiment of the national spirit with patriotism as the core, and a practical model of the socialist core values. The heroic spirit of the people's army is an important spiritual wealth for self-confidence and the development of socialist culture with Chinese characteristics. The moral practices of the people's army have always played an exemplary and leading role in the process of socialist revolution and construction [(Liu and Li, 2020), p. 74]

PLA ethics training encourages the "cultivation of revolutionary soldiers…having [martial] spirit, having [martial] skills, having courage, and having moral character; the so-called "Four Haves" [(Jia, 2017), p. 4]. The term "spirit" is also used throughout PLA ethics writings to explain the desired characteristics that military ethics are to instill in PLA troops. For example, soldiers are encouraged to follow the spiritual examples of selfless soldiers (e.g., Lei Feng Spirit) and even nationwide political campaigns (e.g., The Resist "SARS" Spirit) [(Liu and Li, 2020), p. 73–74].

PLA military ethics also encourage personnel to conform to socialist and traditional Chinese norms, such as collectivism and selflessness. In recent years, this task has been made difficult due to domestic societal changes and perfidious Western influences. The challenges of turning PRC youth who are increasingly enamored of individuality, making money, or their mobile phones into effective PLA soldiers are frequently mentioned [(Tang, 2016), p. 2–4].

The PLA realizes that the need for ethics training extends beyond merely training personnel to obey the Party. Ethics challenges that are created by emerging technologies (such as AI) must also be addressed. Initially established to tackle the unique roles and responsibilities of PRC defense industry personnel, this topic is also used to address the ethical issues faced by soldiers using ever more capable and lethal weapons systems.

> …the relationship between men and weapons are again being developed from a new starting point. The face of warfare is becoming increasingly vague. In modern troop building, military activities, and combat the factor of morality is becoming greater and greater and the matter of military ethics culture is receiving extensive interest. On one hand, the modernization construction of our country's national defense and troops is generating a large number of ethical questions…ethical questions in military training and education, ethical questions in high tech weapons development, ethical questions in military systems, ethical questions regarding military and civilian relationships, knowledge questions regarding the law of war and warfare ethics, etc. They all become questions that must be confronted and settled when reforming a Strong Military. [(Tang, 2016), p. 2–3]

Ethicist Zhao Feng further argues that new and unique ethical issues must be considered as new technologies are developed [(Zhao, 2014), p. 112]. Whether addressing ethical concerns of individuals soldiers or the development of state-of-the-art weapons systems, however, PLA military ethics training consistently emphasizes the incontestable fact that the Party controls the PLA.

PLA military AI ethics

The PLA is intensely involved in applying AI to their capabilities. Taiwan Army Colonel Jing Yuan-Chou explains that the PLA considers AI to be a "'game-changing' critical strategic

1 Metcalf, M. Forthcoming. "A survey of 21st century PLA scholarship on the role of military ethics in warfare," in *Warfare Ethics in Comparative Perspective: China and the West,* eds S. B. Twiss, P.-C. Lo, and S. B. Chan (London: Routledge).

technology; increased machine speed and processing power are expected to be applied to military planning, operational command and decision support as part of the 'intelligentization' of warfare." Xi Jinping has directed the PLA to "accelerate the development of military intelligentization," an endorsement that Jing argues "elevates the concept of intelligentization as a guideline for future Chinese military modernization" (Jing, 2021). The "intelligentization" that Jing describes *specifically* refers to the use of AI to enhance military capabilities and such enhancements result in "intelligentized warfare."

Given this interest, it may seem surprising that seemingly nothing is available from PLA sources regarding *specific* actions that the PLA is considering to address military AI ethical issues. While this can somewhat be attributed to a PLA penchant for security, there are also political factors resulting from Party control of the PLA. For example, when considering the use of LAWS, at a certain point in the process the PLA operator will "relinquish" control of the weapons system to AI functionality; a procedure that is unacceptable in current PLA doctrine. It is ethical questions like this the PLA must address when considering the use of AI.

The PRC and military AI ethics

This does not, however, imply that the PRC is absent from international discussions of military AI ethics. At the Sixth Review Conference of the United Nations' Convention on Certain Conventional Weapons in December 2021, the PRC submitted a position paper on the use of military AI which included the following statement on military AI ethics.

> In terms of law and ethics, countries need to uphold the common values of humanity, put people's well-being front and center, follow the principle of AI for good, and observe national or regional ethical norms in the development, deployment and use of relevant weapon systems. Countries need to ensure that new weapons and their methods or means of warfare comply with international humanitarian law and other applicable international laws, strive to reduce collateral casualties as well as human and property losses, and prevent misuse and malicious use of relevant weapon systems, as well as indiscriminate effects caused by such behaviors. (MFA-PRC, 2021)

Characterized as being "more aspirational than actionable," documents like this provide little insight into how the PLA actually views military AI ethics [(Toner, 2022), p. 255–256].

The PLA and military AI ethics: *PLA Daily*

One type of source that has intermittently shed light on PLA military AI ethics considerations is military newspapers. Vetted by the Party and disseminated throughout the PLA, giving them implicit authority, these sources occasionally provide a forum for regimented discussion of cutting-edge technical or operational issues. For example, a *PLA Daily* article cautions readers about the "ethical black hole of intelligentized warfare"

> In the limited practice of intelligentized warfare, the great changes in the style of warfare have raised a series of ethical issues in warfare. In order to correctly understand and handle the relationship between intelligentized warfare and ethics, and to find a balance between technology and human interaction, these ethical issues need to be examined. [(Wu and Qiao, 2020), p. 7]

The authors highlight several ethical issues that are raised by military AI

- The dangers of a virtual battlefield: "Being in the virtual battlefield for a long time may lead to confusion in the judgment of real values, leading to lack of morality and distortion of the concept of war".
- Inadvertently giving rise to terrorism: While intelligentized weapons systems may improve military operational efficiency and shorten the duration of conflict, "these changes have resulted in a lower threshold for waging war, resulting in frequent violent conflicts, which are contrary to the principles of war ethics" and numb the public to the realities of warfare. In contrast, the side lacking intelligentized systems may have no other recourse than to resort to terrorism in response.
- Attribution of responsibility: "Attribution of responsibility is probably the most criticized ethical issue in intelligentized warfare…Unlike traditional warfare, which can be blamed on specific weapon operators, smart weapons themselves have a certain ability to identify and judge independently. Design flaws, program defects, and operational errors may cause smart weapons to temporarily 'short-circuit', and responsibility comes naturally. Designers, producers, managers, users and supervisors are required to share the responsibility. This transfer of responsibility has greatly increased the difficulty of assigning responsibility after the war. It also leads to another ethical dilemma—diffusion of responsibility." [(Wu and Qiao, 2020), p. 7]

The article concludes by stating "technology is a double-edged sword" and bad experiences are an inevitable consequence of development. Much more research will be required before we can "turn intelligentized technology into a technology that is controlled and beneficial to people." This need to maintain control of AI technology is a frequent theme in PLA writings on the ethical challenges of AI and

is consistent with a desire for strict Party control of weapons systems.

A subsequent *PLA Daily* article approached intelligentized weapons systems from a different perspective. A column entitled "In Future Wars, Will 'Unmanned' Take the Leading Role?" presented different viewpoints on the future implications of unmanned combat [(Liang and Hong, 2021), p. 7]. Liang explained that, throughout history, humans have striven to improve their military capabilities and the intelligentization of weapons systems was yet another step in this process. Eventually, nearly all combat operations would be conducted by unmanned systems and this would result in wars with very few human casualties. Hong rejected this view and argued that human contributions to warfare were essential. From the design of intelligentized systems to the initiation of warfare to the command of combat operations, humans would always be involved.

> People are always the equipment controllers and the active factor to bring equipment advantages into play. The more intelligent the weapons and equipment, the more high-level commanders are needed. Therefore, while the battlefield confrontation may be unmanned, combat control must be manned. (Liang and Hong, 2021)

Hong contends that ethical considerations require that "humans are in charge." Claiming that "military ethics is the moral cornerstone that underpins the modern law of war", the author recounts the numerous civilian casualties inflicted by US drones in Southwest Asia and concludes

> Off-site, non-intuitive, and non-contact implementation of combat operations leads to a lower threshold for war decision-making and a weakening of battlefield moral constraints…Only when humans control the "right to fire" of intelligentized weapons and make unmanned weapons and equipment operate according to human assumptions, can human-machine ethical principles be properly implemented. [(Liang and Hong, 2021), p. 7]

The article concludes by explaining that the two viewpoints highlight the reality that there are still many unanswered questions about intelligentized warfare and that readers should do their best as they work toward developing answers. An interesting aspect of this article is that the two discussants present perspectives that are nearly polar opposites. This would seem to imply that the PLA (and, by extension, the Party) is still wrestling with the operational and ethical implications of incorporating AI into their weapons systems. It is also worth noting that the ethical argument is based on conformity with the law of war and not on any other uniquely-PLA aspects of military ethics. Finally, the article doesn't propose any solutions; only that readers be aware of the challenges and work to solve them.

The PLA and military AI ethics: Academic journals

A different perspective on military AI systems was presented by AI and intelligentized systems specialists at the PLA Academy of Military Science who highlighted potentially problematic technical, ethical, and strategic AI issues. Ethical issues considered were

- Moral crisis: How should machine rules for unmanned vehicles be established when "the power of choice is decided by the algorithm"?
- Military [security] leaks: Extensive use of commercial AI systems may expose PLA personnel data that reveals military vulnerabilities to hostile forces.
- Military law deficiencies: How will accountability be determined when an intelligentized weapons systems mistakenly destroys civilian targets?
- Development of a subjective consciousness: The danger of "the emergence of a super intelligence that can evolve itself and might develop…into machines controlling society or even enslaving humanity."
- The emergence of unmanned forces: The threat of one-sided man-vs-machine warfare.

While some of these concerns seem to be more closely tailored to the PLA's perceptions of military ethics (e.g., data breaches affecting combat readiness), each implied that intelligentized systems might result in independent and/or unanticipated operations [(Cai et al., 2019), p. 71–72].

Analyses of foreign research can also provide an awareness of PLA military AI ethics research priorities. Scientists from the National University of Defense Technology conducted PLA-funded research that *specifically* considered the ethical issues associated with LAWS. Interestingly, all of the article's references were from non-PRC publications [(Zhang and Yang, 2021), p. 47]. The authors explained that the fundamental ethical challenges of LAWS are "the dilemma of 'algorithmic differentiation', the dilemma of military needs and collateral damage, and the responsibility gap caused by the dehumanization of lethal decision-making." They argued

> …the current optimal weapon system is a combination of humans and machines, which not only retains the safety and stability of human judgment, but also takes into account the automation advantages of weapon systems. [(Zhang and Yang, 2021), p. 42]

This "meaningful human control" is consistent with the tenets of PLA military ethics that maintain that the army will fight most effectively while under the direct control of the Party.

Discussion

While investigating applications of AI in modern warfare the PLA is actively considering ethics, but our understanding of their effort is quite limited due to a paucity of publicly-available information. There is, however, sufficient information to draw the following preliminary conclusions.

The CCP wants the PLA to implement military AI

Military AI applications offer the promise of new capabilities that could allow the PLA to surpass the capabilities of current and future adversaries. Ignoring AI would put the PLA and the PRC at a strategic disadvantage.

The PLA is actively investigating the challenges of military AI ethics

While security concerns limit outside access to in-house PLA research, publicly-available materials indicate that PLA analysts are closely monitoring Western military AI ethics research—particularly lessons derived from the Western use of UAVs in Southwest Asia. PLA researchers also understand that military AI will result in significantly challenging ethical considerations and are attempting to resolve such issues.

PLA discussions of military AI ethics are not political

When discussing military AI ethics, *none* of the sources discussed Party considerations. Perhaps this is because Party participation is implicit in military ethics discussions, but the absence of political rhetoric is conspicuous by its absence.

PLA analysis of military AI ethics is highly pragmatic

While PLA authors frequently allude to theoretical aspects of military AI ethics (e.g., accountability, dehumanization, etc.), the general trend of the discussions devolve to the highly practical problem of controlling a system that is, by definition, autonomous. This is an important factor when considering military AI ethics because, from a PLA perspective, appropriate ethical behavior is the logical result of following the Party's guidance.

The PLA needs to resolve the issue of "autonomy or control" for its AI weapons systems

In the near term, the PLA will continue to employ AI to enhance existing military capabilities, but not to implement fully autonomous systems. This does not mean, however, that the PLA is not considering the use of LAWS. The PLA, like all militaries, wants its forces to be equipped with state-of-the-art capabilities and is undoubtedly actively conducting LAWS research and development. Once the PLA is able to solve this challenge to their satisfaction, and in spite of public declarations to the contrary, it would be surprising if they didn't add such cutting-edge capabilities to the PLA's arsenal.

Given the limited availability of relevant data, future insights regarding PLA military AI ethics developments must continue to be meticulously gleaned and interpreted from authoritative PLA journals and official media—particularly since the PLA and Party are apparently still wrestling with such policies and are disinclined to publicly discuss their deliberations. While exchanges between PRC and non-PRC military ethics specialists could provide additional insights, given the current international political climate, prospects for such interactions seem unlikely.

Author contributions

MM conceived of and designed the study, translated and analyzed the Chinese language materials, and wrote all sections of the manuscript.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Cai, C. X., Tian, G. Y., Jia, Z., and Pu, B. (2019). Risk considerations regarding the military use of artificial intelligence technology. *Milit. Operat. Res. Syst. Eng.* 33, 70–73.

Jia, J. (2017). Several suggestions for cultivating a new generation of 'four haves' soldiers in the new era. *Charming China* 18, 7.

Jing, Y. C. (2021). *How Does China Aim to Use AI in Warfare?* The Diplomat. Available online at: https://thediplomat.com/2021/12/how-does-china-aim-to-use-ai-in-warfare/ (accessed October 13, 2022).

Jullien, F. (1995). *Detour and Access: Strategies of Meaning in China and Greece.* Translated by Sophie Hawkes. New York, NY: Zone Books.

Liang, S., and Hong, H. J. (2021). *In Future Wars Will, "Unmanned" Take the Leading Role?* PLA Daily, 7. Available online at: http://www.81.cn/jfjbmap/content/2021-03/30/content_286005.htm (accessed October 13, 2022).

Liu, S. P., and Li, L. (2020). The strategic value of the development of contemporary military ethics. *J. Nanj. Univ. Sci. Technol.* 33, 72–76. doi: 10.19847/j.ISSN1008-2646.2020.04.012

MFA-PRC (2021). *Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI).* Ministry of Foreign Affairs of the People's Republic of China. Available online at: https://www.fmprc.gov.cn/mfa_eng/wjdt_665385/wjzcs/202112/t20211214_10469512.html (accessed October 13, 2022).

Tang, F. (2016). *An Investigation of Building a Contemporary Chinese Military Ethics Culture.* Shanghai: Shanghai World Publishing Company.

Toner, H. (2022). "AI safeguards: views inside and outside China," in *Chinese Power and Artificial Intelligence*, eds W. C. Hannas, and H.-M. Chang (London: Routledge), 244–259.

Wu, J. X., and Qiao, P. (2020). *Pay Attention to the Black Hole of Intelligentized Warfare.* PLA Daily, 7. Available online at: http://www.81.cn/2020fdqj/2020-04/28/content_9842086.htm (accessed October 13, 2022).

Zhang, Q., and Yang, A. H. (2021). Ethical challenges and risk responses for lethal autonomous weapon systems. *Stud. Dialect. Nat.* 37, 42–47. doi: 10.19484/j.cnki.1000-8934.2021.03.008

Zhao, F. (2014). On the development of innovation in military science and technology ethics research conforming to military combat readiness. *J. PLA Nanjing Inst. Polit.* 30, 112–118.

frontiers | Frontiers in Big Data

TYPE Perspective
PUBLISHED 06 December 2022
DOI 10.3389/fdata.2022.1038507

OPEN ACCESS

EDITED BY
Henrik Syse,
Peace Research Institute Oslo (PRIO), Norway

REVIEWED BY
Paul Lushenko,
Cornell University, United States

*CORRESPONDENCE
Patrick Taylor Smith
patrick.taylor.smith@gmail.com

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 07 September 2022
ACCEPTED 16 November 2022
PUBLISHED 06 December 2022

CITATION
Smith PT (2022) Resolving responsibility gaps for lethal autonomous weapon systems. *Front. Big Data* 5:1038507. doi: 10.3389/fdata.2022.1038507

Resolving responsibility gaps for lethal autonomous weapon systems

Patrick Taylor Smith*

Stockdale Center for Ethical Leadership, United States Naval Academy, Annapolis, MD, United States

This paper offers a novel understanding of collective responsibility for AI outcomes that can help resolve the "problem of many hands" and "responsibility gaps" when it comes to AI failure, especially in the context of lethal autonomous weapon systems.

KEYWORDS

lethal autonomous weapon systems, ethics of artificial intelligence, military ethics, responsibility, responsibility gaps, ethics of technology

Introduction

This paper provides the normative grounding and a general description of a political conception of responsibility for just war compliance and non-compliance by lethal autonomous weapon systems. Deploying the Unfair Burden Argument, the Agent Constitution Argument, and the Collective Values Argument, the paper shows that we should move away from an interpersonal and ethical understanding of responsibility to a collective and holistic distributive conception of responsibility where we assign various accountability mechanisms and responsibilities to agents in the system on the basis of effectiveness and fairness rather than direct moral responsibility. This new account dissolves the problem of responsibility gaps for lethal autonomous weapon systems and points a way forward toward appropriately and legitimately distributing responsibility through the defense statecraft ecosystem.

Research article

The problem of responsibility gaps for lethal autonomous weapon systems

Lethal autonomous weapon systems (henceforth LAWS) are a set of proposed and speculative systems—though increasingly plausible—that mediate between human agency and the use of lethal force. Unlike automated systems that can fire on their own, autonomous weapon systems have the capability to function independently in a chaotic battlespace with little proactive human intervention (Horowitz and Sharre, 2015), operating along the kill chain without full human supervision. That is, they can use their own sensors and algorithms to acquire their targets and "decide" to fire on their own

without a human decision. Such systems are obviously controversial and many militaries have rejected their use, but they have considerable potential utility, especially in cases where human reactions will be too slow or where communications between operator and drone are disrupted. Since LAWS are meant to operate on their own, they can operate unpredictably and act in ways that no individual operator or programmer would endorse. We can imagine cases where every human agent does what they can to reasonably foresee potential failure points and the unpredictable nature of the interaction between system and environment leads the LAWS to engage in an indiscriminate or disproportionate attack that violate the standard strictures of *jus in bello*. These actions could potentially involve the deliberate killing of non-combatant or the use of disproportionate force in a way that is unjust and immoral (Walzer, 2000).

As a consequence of this chaotic and unpredictable autonomy and the corresponding likelihood of just war violations, LAWS will almost certainly be involved in "responsibility gaps" (Sparrow, 2007; Asaro, 2012; Santoni and van den Hoven, 2018) where the system does something immoral and yet no person can be held accountable. Thus, LAWS might problematically "off-shore" potential responsibility by having the LAWs "make decisions" where it is genuinely unclear if anyone is truly responsible for the violation. There appears to be something uniquely bad about a non-human entity "making the decision" to violate the requirements of just war without any possibility of holding the violator responsible for the violation. Thus, the ultimate permissibility or impermissibility of LAWS may depend on finding a way to resolve the responsibility gaps inherent in its operation. Of course, there may be other reasons for rejecting unmanned, autonomous weapons (Emery and Brunstetter, 2015), but this paper is focused on evaluating and responding to the concerns generated by responsibility gaps.

Nonetheless, the notion of "responsibility" whereby the LAWS generates responsibility gaps is individualist (Smith, 2018). On this view, we need to develop an elaborate account of interpersonal ethics that tells us exactly the nature of our moral contribution and our level of responsibility for a particular set of decisions or consequences. Our obligations are then derived from this rigorous understanding of our responsibility. This generates the problem of responsibility gaps since it is not obvious that any one person can be assigned the relevant interpersonal moral status. But as we shall see, there is a different and more political way to understand responsibility: agents are responsible for what they would agree to under fair decision-making conditions (Simpson and Mullers, 2016). So, this paper takes the responsibility gap problem seriously but argues that it can be resolved if we understand responsibility differently in the military context.

Rawlsian institutionalism and distributed responsibility

In what follows, I argue that we should adopt a "division of labor" whereby institutions *assign* individual responsibility much like institutions provide individuals with distributive shares. On this view, a person is responsible when an institutional *fairly* ascribes responsibility to the agent. Following a broadly Rawlsian understanding of the institutional division of labor in distributive justice (Rawls, 1971, 2001, and Pogge, 2000) an institutional division of labor is justified under three conditions (Smith, 2022). I will explain this idea in greater detail later, but first I will justify *why* we should move to a more collective understanding of responsibility in the first place. Thus, distributed responsibility is not a replacement for individual responsibility, but rather a supplement to it when particular conditions obtain. It is a matter of focus: do we start with institutions and *derive* individual obligations, or do we start with individual obligations and treat institutions as instruments for meeting them? Often, we do the latter, but there are conditions when we should do the former. This institutional priority and focus is also what distinguishes my view from other "distributed responsibility" views (Galliot, 2015) that nonetheless still start with the individualist, interpersonal conception.

First, a division of labor between individual and institutional obligations can be justified when satisfying the principles of justice requires practical coordination or epistemic demands that are unreasonable or impossible for individuals acting unilaterally. Taking Rawls's (2001) example, imagine a Lockean understanding of distributive justice of the following kind: we begin from a position of rough equality and then engage in a series of voluntary transactions that are just when they leave "enough and as good" for others. Initially, it might be possible for each person to have sufficient information and be able to anticipate what others are doing such that they could be reasonably confident that their individual choices satisfied the view. Yet, in any sort of complex society, the informational and cognitive requirements of understanding whether one was leaving "enough and as good" would be enormous. It would be unreasonable to expect any particular agent to be able to reliably make those judgments in all distributively relevant contexts. Similarly, ensuring that each person has the resources to engage in the foregrounded voluntary transactions will require intense practical coordination in terms of how much to give, what to give, and who should give. The idea here is that our obligations are *entangled* and that there is no *a priori* answer between various coordinative equilibria. As a consequence, there simply is no correct answer about the appropriate individual obligations without some authoritative, coordinative mechanism to determine individual contributions. And even if there were an optimal equilibrium to be discovered, this would only add to the informational and

calculative burdens of individual agents. So, institutionalism can be justified when individual satisfaction of the principles is made impossible, unfair, or unreasonable by informational or coordinative burdens. In other words, institutions—by which I mean structures that use general rules and norms to purposively coordinate and direct human behavior—are required in order to maintain the background conditions for individuals to make fair and voluntary choices in their day-to-day economic, social, and political interactions. Call this the *unfair burden argument* for institutionalism.

Yet, we still might claim that the institution should be trying to replicate the *ideal*, aggregative choices of individual agents rather than claim that individuals have fundamentally different obligations from institutions. If we thought that imposing ideal individual obligations on actual individuals was unfairly or unreasonably burdensome, we would still want a moral division of labor, but we might still think that institutionalism was just there to "help" individuals satisfy their individual obligations. But there is another set of reasons for an institutionalist focus. If institutions play an essential role in *creating* and *maintaining* the agential capacities, powers, and resources that make it possible for individuals to propose, discuss, and abide by reasonable principles of justice, then we would need principles of justice that apply to those institutions over and above that of individuals. Insofar as institutions play an essential role in constituting the agency of the individual actor and have a large influence over the choice structure presented to the agent, then principles of justice need to apply to the institutions themselves. Otherwise, we will be imposing obligations upon agents without understanding or regulating the core influences upon that agent. It would seem odd to argue that individuals need to bear considerable burdens when faced with certain choices and not normatively evaluate the profound influence that the government, the family, or the market has over whether and to what extent the agent will have the capacities or resources to engage with those requirements in the first place. Call this the *agential influence argument* for institutionalism.

The *agential influence argument* provides an indicator of when institutionalism is necessary: different institutions will produce different agents with different capacities, facing different choices and circumstances. The *unfair burden argument*, on the other hand, suggests that we assign distinctive responsibilities to institutions. Combined, they suggest a kind of moral primacy for institutions for at least *some* questions: being a virtuous agent will do little to guarantee compliance with the relevant principles and good institutions can permit agents to be more self-interested and still produce just outcomes. So, if these arguments apply to a normative domain, then we have good reason to adopt an institutionalist paradigm whereby institutions are regulated by the principles of justice and individuals have an obligation to support those institutions and follow their dictates.

Finally, institutionalism may be justified when there are distinctive political values that can only be expressed or instantiated by collective institutions. For example, if deliberative democracy makes it possible for citizens to engage in binding, collective decision-making and it is an important political value that I participate in decisions that affect my core interests, we might think that the institutions of democratic decision-making are necessary for everyone to engage in legitimate, coordinated action. Similarly, Kantians and neo-republicans (Pettit, 1997; Young, 2000; Stilz, 2011) both argue, though for somewhat different reasons, that rightful relations between persons can only be achieved if mediated through political institutions that provide guarantees of their freedom from the domination of others. However, since this freedom needs to be assured by something other than the individual virtue, it is impossible for an individual to bring about these values on their own. If we accept these accounts of political freedom, then we must be institutionalists about—at least—these values as it is only through institutions that they are possible. Call this the *collective values argument*.

Applying Rawlsian distributed responsibility to LAWS

In this section, I do two things. First, I show that these three arguments for institutionalism apply to lethal autonomous weapon systems. Second, I then show how institutionalism might be applied to resolve responsibility gaps for LAWS.

Let's take each of the three main arguments in turn. First, *unfair burden*. The chaotic and unpredictable nature of AI driven technology, even when well-tested validated, combines with the chaotic and unpredictable nature of the battlespace to make it very difficult, if not impossible, for individuals to make reliable, effective judgments with enough speed to prevent just war non-compliance. The cognitive burden of managing drone-human teams under chaotic conditions and the consequent unfairness of applying full responsibility to the user or commander is one of the drivers of responsibility gaps in the first place. For example, imagine that a commander is operating a "centaur" human-drone hybrid where the drone uses an algorithm to determine whether a target is a lawful combatant. The drone is in the process of "clearing" a room and determining it is safe for humans to enter and makes a split-second judgment that a person in the room is a combatant and kills them. It is very unlikely that the commander of the drone, or any member of the team, will always be able to intervene in real time to evaluate whether the drone is correct and then intervene to stop it if it is mistake. First, the drone is using perceptual capacities—radar, lidar, and the like—that the commander cannot easily process and is using rapid calculations to aggregate that data much

faster than a human can comprehend. Even if the algorithm was explainable, the process would go by too quickly for the commander to remain "in the loop." It is unreasonable to expect them to be able to do so. Thus, it seems plausible—as others have argued (Hayry, 2020; Verdiesen et al., 2021), but without the political foundations of this piece—that we need a broader understanding of institutional responsibility in the face of these concerns.

Second, LAWS will invoke concerns about agential constitution because these technologies will shape the very agency of the humans who will be participating in human warfighting. First, they will affect perception as the autonomous drones will feed information back to human warfighters, perhaps in spectrums and in formats that humans themselves cannot even perceive. Thus, drones will become part of our agency just as eyeglasses and hearing aids have become part of our agency, and this trend will only increase as we develop close-knit centaur human-AI teams as humans will be able to "see" the battlefield in certain ways due to their drone counterparts. Further, humans will come to understand what they can "do" in terms of delivering fire and shaping the environment in terms of what their drones can do. A human commander will understand that "they" can clear a room without using deeply coercive measures using LAWS but will also come to feel as they have decreased capacity when those drones are unavailable, just as we feel a reduction in our own capacities once the wireless internet stops working. In other words, we shape our own capacities based on the expectation that tools and technologies will be able to take up the slack, such as we when we stop memorizing phone numbers because smart phones will store them for us. That means, our own cognitive and physical capacities are structured by what we expect our tools to be able to do. A focus on individual moral responsibility at the cost of institutional distributed responsibility will miss the ways doctrinal, design, and deployment choices will shape the vary ways that humans act and perceive.

Finally, there are a plethora of collective and political values that apply to military action. Just war theory—as well as international law—is structured by the normative demand for proper authority: so appropriate constitutional legitimacy is a key a feature of the right to go to war (Fabre, 2008; Galliot, 2015). The use of autonomous systems in the military context will require *trust*, which is a feature of the institutions themselves. When an individual warfighter uses a drone, they are trusting a complex set of institutions that engaged in design, testing, and validation and whether those processes are trustworthy is a collective value. Also, protecting the rights of non-combatants and civilians who are subject to the authority and coercive power of soldiers requires more than just that soldiers *individually* refrain from war crimes, the rights of civilians must also be *assured* by substantial accountability mechanisms that mitigate the arbitrary authority and power that military personnel can have over civilians. Finally, some have argued that the practice of atoning for military ethics violations must be collective as individual soldiers will not be able to go through the practice of apology and reparations for individual victims. In general, soldiers operate within a collective context where what they do reflects on the collectivity and what the collectivity does reflects upon them, both for good and ill. Many of these considerations apply to the military in general, but these issues are only exacerbated with LAWS.

So, let's grant that we need an institutionalist orientation for responsibility for LAWS compliance and non-compliance just war principles rather than an interpersonal one. It would take too much space to fully delineate how this would work in practice, but I will offer some preliminary comments. There are three elements of a Rawlsian distributed responsibility account: an account of what is to be distributed, an account of the institutions that work together to distribute the responsibility and produce the normatively relevant outcomes, and a process to choose fair principles of distribution. Let's take each in turn.

First, the account concerns the distribution of responsibility, but it is essential to see that we can pull apart the various ways we hold people accountable. We hold people responsible in many ways: criminal liability, civil liability, career-oriented costs and benefits, and social opprobrium, amongst others. There is no reason that a political system would distribute these various mechanisms uniformly; instead, we should *disaggregate accountability mechanisms*. Suppose we believe that both a LAWS designer and a commander who deploys a LAWS that violates the principles of just war should be held responsible for the failure. On an interpersonal view, we might think the question is "responsible or not?" but on the political view the question now becomes, "What *sorts* of responsibility should we distribute onto the various agents?" So, we might hold the corporation who designed the LAWS civilly liable to compensate the victims while holding the commander liable through the diminution of their career prospects while saving criminal liability for other agents and social opprobrium for yet others. Again, accountability mechanisms are disaggregated and then distributed throughout the system to produce good outcomes in a fair way. This is one way that my more political conception is different from other collective responsibility views: they treat "responsibility" as a monolithic notion rather than one that can be disaggregated.

Yet, how should accountability be distributed such that it is fair? A final determination is beyond the scope of this paper, but I would like to describe how a broadly Rawlsian-constructivist (James, 2005) account might proceed based on the *veil of ignorance*. First, we would understand the complex set of institutions that produce LAWS outcomes as a kind of cooperative endeavor: political oversight. design, testing, evaluation, validation, training, doctrine, and deployment all work together as a web-like system of systems to generate a contextual rate of just war compliance by the specific LAWS that is created and used. A Rawlsian-constructivist—not necessarily

Rawls himself—understanding of this cooperative structure lends itself to the following question, "Given the need to generate the relevant ethical values, how would we distribute various accountability mechanisms if we did not know where in the cooperative system we might find ourselves?" In other words, who would we hold accountable and why if we were ignorant of how those decisions might come to apply to us? This is a way of modeling fair decision-making as it prevents one from biasing the distribution based upon their knowledge that they will be powerful agents in the system and focuses attention on the common good (Huang et al., 2019). If I knew I was going to be a high-ranking officer, politician, or corporate executive, I might design a system that shields me from accountability. Yet, this is far less attractive if I do not know if I will be the executive or a young lieutenant facing the decision to use the drone in combat; the veil of ignorance forces me to decide on principles and distributions for everyone on an equal basis because I could be anyone in the system.

A consequence of these two features—disaggregation and the veil of ignorance—of institutional responsibility is that accountability will be distributed far more widely and holistically than one might traditionally believe and that there should be consequences for failure up and down the chain of decision-making for LAWS outcomes. If I knew I might be a young lieutenant deciding whether to deploy LAWS and that I would be held at least partially accountable for what happens, then I would demand principles that assigned accountability to other agents to ensure that I was placed in a position to succeed and that I could trust the reliability of the system. So, responsibility would move beyond the military chain of command to include the civilian leadership making decisions on where to go war and why, the technology and defense contractors designing the system, and the defense bureaucracy making choices on training and doctrine. Of course, if one knew that there was the possibility of being held accountable for the choices of the tactical commander in the field, then a system where the tactical commander had no responsibility for what happens would also be unacceptable. What is needed is to balance the relevant claims of the stakeholders within the defense statecraft ecosystem and for that, we need a political conception of distributed responsibility.

I will end this paper with a brief anecdote. I have taught military ethics to both experienced officers and midshipmen still waiting on their commissions, and they are taught to take responsibility to prevent war crimes and atrocities. Yet, I have also been shown the computer simulations used by defense consultants to wargame tactical decision-making and, indirectly, to contribute to doctrine and procurement. The very tools my midshipmen will possess are, in part, determined by these simulations. Yet, these simulations include *no* provision for preventing civilian casualties; it is not that they are ignored, it is that civilians do not exist. The consultants take essentially no responsibility in ensuring that warfighters have the appropriate tools to achieve their objective within the context of the rules of war as just war principles are left to others. This is both unsurprising and perfectly rational in the context of the interpersonal model: their contribution is far too indirect to activate individual, personal responsibility. Yet, it is deeply unfair that individuals who are much more powerful and well-connected, who have the time and money to think carefully, are "off the hook" while the newly-minted lieutenant facing combat for the first time feels the full brunt of accountability. Of course, military officers receive special training and develop specific virtues to handle this sort stress and this is relevant to responsibility attributions, but having power and authority within the system is *also* relevant. And this is especially true when the battlespace becomes populated by objects as complex as LAWS. To resolve this problem, we must reorient our thinking in a political direction.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This work was funded by Peace Research Institute Oslo, Norway for the Open Source fee.

Acknowledgments

I would like to the participants of the Stockdale Research Fellow Seminar and the audience at the Rocky Mountain Ethics Congress for their helpful comments. I would also like to thank Ryan Jenkins for his detailed commentary on an earlier draft.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *Int. Rev. Red Cross* 94, 687–709. doi: 10.1017/S1816383112000768

Emery, J., and Brunstetter, D. (2015). Drones as aerial occupation. *Peace Rev.* 27, 424–431. doi: 10.1080/10402659.2015.1094319

Fabre, C. (2008). Cosmopolitanism, just war theory and legitimate authority. *Ethics and Int. Aff.* 84, 963–976. doi: 10.1111/j.1468-2346.2008.00749.x

Galliot, J. (2015). *Military Robots: Mapping the Moral Landscape*. Surrey: Ashgate Publishing. doi: 10.4324/9781315595443

Hayry, M. (2020). Employing Lethal autonomous weapon systems. *Int. J. Appl. Philos.* 34, 173–181. doi: 10.5840/ijap2021326145

Horowitz, M., and Sharre, P. (2015). "Meaningful human control in weapon systems: a primer," in *Center for New American Century Working Paper* (Project Ethical Autonomy).

Huang, K., Greene, J., and Bazerman, M. (2019). Veil of ignorance reasoning favors the greater good. *Proc. Natl. Acad. Sci.* 116, 23989–23995. doi: 10.1073/pnas.1910125116

James, A. (2005). Constructing justice for existing practice: Rawls and the status Quo. *Philos Public Aff.* 33, 281–316. doi: 10.1111/j.1088-4963.2005.00034.x

Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Princeton, NJ: Princeton University Press

Pogge, T. (2000). On the site of distributive justice: reflections on Cohen and Murphy. *Philos. Public Aff.* 29, 137–169. doi: 10.1111/j.1088-4963.2000.00137.x

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674042605

Rawls, J. (2001). *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University of Press

Santoni, S. F., and van den Hoven, J. (2018). Meaningful human control over autonomous systems: a philosophical account. *Front. AI Robot.* 5. doi: 10.3389/frobt.2018.00015

Simpson, T., and Mullers, V. (2016). Just war and robots' killings. *Philos. Quart.* 66, 302–322. doi: 10.1093/pq/pqv075

Smith, P. T. (2018). Just research into killer robots. *Ethic. Inf. Technol.* 21, 281–293. doi: 10.1007/s10676-018-9472-6

Smith, P. T. (2022) "Distributive justice, institutionalism, and autonomous vehicles," in *Autonomous Vehicle Ethics: The Trolley Problem and Beyond, eds* R. Jenkins, et al. (Oxford: Oxford University Press).

Sparrow, R. (2007). Killer robots. *J. Appl. Philos.* 24, 62–77. doi: 10.1111/j.1468-5930.2007.00346.x

Stilz, A. (2011). *Liberal Loyalty: Freedom, Obligation, and the State*. Princeton, NJ: Princeton University Press.

Verdiesen, I., Santoni de Sio, F., and Dignum, V. (2021). Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. *Minds Mach.* 31, 137-163. doi: 10.1007/s11023-020-09532-9

Walzer, M. (2000). *Just and Unjust Wars. 3rd edn.* New York, NY: Basic Books.

Young, I. M. (2000). *Justice and the Politics of Difference*. Princeton, NJ: Princeton University Press.

frontiers | Frontiers in Big Data

TYPE Perspective
PUBLISHED 09 January 2023
DOI 10.3389/fdata.2022.1017677

OPEN ACCESS

EDITED BY
Edward Barrett,
United States Naval Academy,
United States

REVIEWED BY
Patrick Lin,
California Polytechnic State University,
United States
Charles Pfaff,
United States Army War College,
United States

*CORRESPONDENCE
Jovana Davidovic
✉ jovana-davidovic@uiowa.edu

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 12 August 2022
ACCEPTED 21 December 2022
PUBLISHED 09 January 2023

CITATION
Davidovic J (2023) On the purpose of meaningful human control of AI. *Front. Big Data* 5:1017677. doi: 10.3389/fdata.2022.1017677

On the purpose of meaningful human control of AI

Jovana Davidovic[1,2]*

[1]Department of Philosophy, The University of Iowa, Iowa City, IA, United States, [2]Stockdale Center for Ethical Leadership, United States Naval Academy, Annapolis, MD, United States

Meaningful human control over AI is exalted as a key tool for assuring safety, dignity, and responsibility for AI and automated decision-systems. It is a central topic especially in fields that deal with the use of AI for decisions that could cause significant harm, like AI-enabled weapons systems. This paper argues that discussions regarding meaningful human control commonly fail to identify the purpose behind the call for meaningful human control and that stating that purpose is a necessary step in deciding how best to institutionalize meaningful human control over AI. The paper identifies 5 common purposes for human control and sketches how different purpose translate into different institutional design.

KEYWORDS

machine learning and AI, war, meaningful human control, ethics, robots

Introduction

All around us algorithms are making decisions about us and for us. From how we chose what to watch, to how we shop, get policed and go to war, get social services, medical diagnoses, or loans, algorithms are quite literally everywhere. No aspect of our lives is unaffected by algorithms whose incredible power promises to continue to change our lives. AI, big data, and machine learning will help us address climate change, cure cancer, feed more people, and fight less bloody wars. But with this incredible power comes great potential for harm. In fact, the very features that make AI a powerful tool also make it very dangerous. These features include the ability to process large data sets that humans cannot, the ability to "see" patterns humans could not, the ability to apply solutions on grand scales, and the ability to do so at great speed. These abilities and their driving force—machine learning—make AI not only capable of causing harm, but also less transparent, less explainable, and often unfair and unjust. Much has been said about these issues (Mittelstadt et al., 2016; Felzmann and Villaronga, 2019; Larsson and Heintz, 2020; Brown et al., 2021). Lawyers, scholars, and the public have, for example, repeatedly called for transparency and explainability, arguing that we cannot leave morally consequential decisions to machines. Instead, they argue, we need human-machine teams, and the necessary transparency and explainability for those teams to work.

Most importantly, many scholars argue, we need meaningful human control over these powerful algorithms.[1] After all, we need a human to stop a drone attack on a person identified by a classification algorithm as a combatant, but carrying a carpet rather than a rocket launcher. We need a human to question or re-assess a recidivism risk assessment when the algorithm is known to be racially biased and intentionally designed to err toward false positives. We want a human to question the predictions of a climate model not trained on the data relevant to geographic location within which we are trying to apply it. Simply put, we need to make sure that the algorithms we use do not cause more harm than they can spare us and to do so we need meaningful human control (UNIDIR, 2014). For example, the U.S. military has called for meaningful human control over certain systems or what they refer to in DoD directive 3,000.09—"appropriate human judgement" (Department of Defense, 2016). Similarly, the U.N. Committee for the Convention on Certain Conventional Weapons (CCW) has argued that in spite of the fact that "there is not yet an internationally agreed definition of what precisely meaningful human control constitutes, there is…convergence that some degree of human control over…LAWS is vital" (Schwartz, 2018). The European Commission also recently proposed a regulation that stipulates that ""high-risk A.I. systems" (such as facial recognition and algorithms that determine eligibility for public benefits) should be designed to allow for oversight by humans who will be tasked with preventing or minimizing risks" (Green and Kak, 2021). The GDPR (European General Data Protection Regulation) also assures the right not to be subjected to a decision based on solely automated process (European Union Parliament, 2016). Examples go on, but the main takeaway is that the primary effort to mitigate risks of harm from ADS (automated decision-systems) and AI across jurisdictions focuses on assuring "meaningful human control." The problems with this approach are many—from the fact that scholars do not agree on what meaningful human control is, to automation bias, i.e., the tendency to trust machines when machines and humans opinions conflict, to the worries that we cannot expect humans to meaningfully provide oversight for the very systems that were built because of and for things humans do not have the capacity to do (Green and Kak, 2021). These are significant problems for sure, and solving them will be key for mitigating the risks of ADS and machine-learning AI. But here I want to focus on what I see as the *conceptually primary problem*—clarifying the purpose of meaningful human control.

The success of meaningful human control as a "solution" for the woes of AI depends on the problem one is trying to solve for. Generic calls for meaningful human control are unhelpful and have consistently led to generic descriptions of what meaningful human control would look like. Discussions of "meaningful human control" most often focus on who should and when exert control over the process, without ever explicitly asking, for which purpose or why (Roff and Moyes, 2016; Ekelhof, 2018). Simply put, meaningful human control can solve different problems and serve different purposes and as such it requires different institutional design for different aims. Thus, the first step in deciding whether we need meaningful human control, and what shape that human control ought to have, as well as what do we need to successfully exert such control (e.g., what type of explanations or information) depends on the purpose of that control. In what follows, I lay out the 5 main purposes that human control of automated decision systems could serve, and then I concisely explain why and how different purposes require different institutional design and types of different explanations of ADS outputs.

Purpose axis—The five purposes of meaningful human control

a. Safety and precision: One, common, reason for human control over AI systems is accuracy, safety, and precision. In many cases, the reason we hope to have a human in the loop is because we think that that will prevent mistakes and avoid harm. Such calls for "humans in the loop" make sense in cases when humans are better at some cognitive task (object recognition—for now), or when context affects outcomes and is difficult to model, or in cases when unanticipated changes to our environments might occur. In cases when a human together with a machine performs better than a machine alone, safety and precision are an obvious reason to have meaningful human control. Of course, such control might not be possible in cases where large sets of data are processed by the algorithm or when the speed of processing or the need for speed of decision-making is what makes the AI particularly valuable (e.g., anti-missile or anti-drone swarm ship defense systems). Centrally, when the aim of human control is safety, the location of the human in the loop in the decision-making chain, should obviously be driven by increase in safety and precision. Whether the human should be the final decider, or just an oversee-er, or only have control over deployment more generally, when safety is primary concern, should be solely driven by empirical analysis- whatever works more effectively.

b. Responsibility and accountability: Sometimes, meaningful human control is, however, primarily meant to assure accountability and responsibility. In as much as machine learning algorithms or semi-autonomous or autonomous

1 What we mean by "meaningful human control" is an open question. In the narrowest of senses, it means having a person who presses a button somewhere during each use of the AI in question. In the broadest of senses, it means oversight of processes-either in use of AI or in production, acquisition and use of AI.

AI play a role in decisions that might lead to lethal harm or other types of significant harm, institutions using such AI, might be interested in knowing who to hold responsible for potential failures and resultant harm. Where responsibility chains are already prescribed, one might be interested in knowing how to adjust those responsibility chains in cases when a decision relies on AI. We might, for example, ask how to distribute responsibility between developers, acquisition teams, and those that choose to deploy the system in a particular setting. If our primary concern is assigning responsibility, we might "insert" a human in a different part of the algorithms' life cycle then we would have if our primary concern is safety. For example, unlike the cases when our primary concern is safety, in cases where we want meaningful human control for purposes of responsibility, we might take into consideration previous responsibility assignments, or even arbitrary assignments of human control (as long as they are clear).

It is worth nothing that responsibility assignment and accountability might not require the same solutions and are not identical. Accountability, in some cases, simply requires that we know why the decision was A rather than B (for example so we can assess whether the reasons used for a decision were constitutional, or fair, or reasonable). Accountability might, therefore, at times, be satisfied by a simple technological solution. For example, a meta-interpretive algorithm like LIME (Local Interpretable Model Agnostic Interpretations) (Ribeiro, 2016). Responsibility assignments cannot, in contrast, be satisfied technologically. Responsibility, at least for now, requires a human in the loop for different reasons—because as it stands we can't hold machines responsible in any meaningful sense.

In addition to the fact that the shape and location of human control for purposes of accountability and for purposes of responsibility vary, it is also important to note that there is a range of types of responsibility-purposes. For example, assignments of moral responsibility and assignments of legal responsibility might require different types of institutional design for "meaningful" control. When assignments of responsibility are the reason behind the calls for meaningful human control, it matters greatly whether we are after:

b. i. Legal responsibility.

1. Forward-looking (for which corporate liability models might act as a potential model) (Elish, 2019; Selbst, 2020; Diamantis, 2021).
2. Backward-looking (for retributive or restorative justice).

b. ii. Moral responsibility.

1. Moral responsibility for assigning blameworthiness.
2. Moral responsibility for assigning liability to defensive harm.
3. Moral responsibility for assigning liability to punitive harm.

c. Morality and dignity: Another common reason people have called for meaningful human oversight is to solve for problems they see with harm and especially lethal harm being imposed by fully autonomous weapons systems, sometimes called "killer robots" (Horowitz and Scharre, 2016). Those arguing against killer robots usually argue that fully autonomous AI doesn't have key moral features (moral reasoning for example) and thus meaningful human control is needed to justify lethal harm (Purves et al., 2015). Others argue that to be killed by a machine violates human dignity and thus a human is needed in the loop any time lethal harm is considered. Meaningful human control for purposes of assuring dignity of targets will obviously take a very different form than meaningful human control for purposes of, for example, legal responsibility. For example, while legal responsibility can be captured by some kind of strict liability approach—in which case owner of the ADS would be the one considered in "control" and thus responsible for its malfunction, dignity on most accounts requires that a human is the final link in the kill chain, and in a meaningful sense—the "proximate cause" of one's death.

Of course, issues of this kind also exist outside of warfighting contexts—there might be dignity-related reasons to want human control over, for example, biomedical decisions—like end of life decisions, or over the distribution of medical resources, or social services. One might argue for example that there is something morally problematic with leaving medical decisions to ADS without a human in the loop even when safety and precision are not at stake.

d. Democratic engagement AND consent: Often, human control and engagement, have little to do with, or are only instrumentally related to, lowering harm and increasing safety, but instead, are required for procedural justice and fairness. Sometimes we might want stakeholders or those to whom the algorithm is applied to, to have sufficient understanding of the process to consent or dissent and in that way provide human oversight and control over the algorithm (Brennan-Marquez and Henderson, 2019; Pasquale, 2021). In these cases, the benefit of the control is primarily aimed at either democratic engagement or justified consent, and in these cases, the institutional shape that meaningful human control will take will obviously be quite different from cases where it is meant to simply or

solely minimize harm. For example, we might be interested in human control over parole decisions, not only to have a recidivism risk tool that is precise and has equal and small false positive rates across racial groups, but we might also want enough transparency in such algorithms so that those to whom the algorithm is applied can challenge specific assessments/outputs of the algorithms as they apply to them. Similar arguments can be given for transparency and explainability for any juridical ADS—namely that one shape meaningful human control can take is ability to question the decisions by such ADS.

e. Instituional stability: There might also be times when the benefit of meaningful human control is really only in the appearance of such control- this might have to do with cases when we want to provide reasons for trust in the institution (Brennan-Marquez et al., 2019). If we are solely after the appearance of meaningful human control, such "control" might look very different, then if we are after the control for one of the above reasons. There might for example be times where appearance of meaningful human control is simply the best we can do, but as a matter of institutional success and stability, such appearance of human control is helpful. Arguably, some autonomous vehicle systems might still rely on having a human on the loop (as a back-up) even if and when that doesn't statistically alter safety, if it increases the trust of pedestrians and society. Whether or not these are good reasons to have meaningful human control, is less relevant here, what matters is that when this is the (or a) reason for such control, it should drive the institutional design around "meaningful control."

It should be noted that more than one of the purposes discussed here could be behind any particular call for meaningful human control, but being explicit about the main purposes and understanding the institutional design that would best serve each purpose is a crucial first step in trying to make the changes so many are calling for.

Let me finally say a bit more about what it means to say that knowing the purpose of meaningful human control drives institutional design. Understanding the purpose behind calls for "meaningful human control" will provide: (a) the building blocks for the type of explanations we might need and (b) the appropriate location for the meaningful human control. In fact this is the primary reason we should care about carefully and explicitly stating the purpose behind a call for human control over some automated decision system.

Regarding explainability, explainable AI is needed, scholars argue, to be able to exert meaningful control, to be able to justify our actions to citizens, and to be able to question and challenge an ADS decision (Alan Turing Institute, 2020). Scholars often follow up calls for explainable AI (XAI), by lamenting that fully explainable AI is not possible and thus we are stuck with all the problems or many of the problems of ADS (Newman, 2021). But it matters greatly what kind of explanations we are after. We do not always need full explainability, and type of control we are after drives the type of explanation we are after. There is an abundance of literature on explainable AI and many techniques are being developed to apply to (for example classification) algorithms. Developers are opting for more explainable methods more often. Knowing why we need human control and at which stage drives the shape we want XAI to take in a particular setting. Explicit statement of purpose of meaningful human control will thus not only help shape institutional design around the algorithm, but also the shape our explanations need to take to satisfy that purpose. And thus, knowing the purpose of human control, will allow us to be more precise in asking for explainable AI. For example, for some end users whose primary focus in safety it is sufficient that they know common ways a system might fail- and really the only "explanation" they need is to know when not to trust the system. Others who might need to exert control over an algorithm might need to be provided explanations regarding training data- so as to be able to anticipate when a system might not perform well in a new environment. In cases when our primary reason for a call for meaningful human control is responsibility assignment, we probably want the person responsible to have enough of an explanation to be able to form a justified belief—otherwise they may never justifiably use an algorithm and on some accounts of responsibility might never be responsible for negative outcomes, since they wouldn't be held responsible for their ignorance.

Similarly, knowing and explicitly stating the purpose of human control will drive the location where such control is best exerted. If we think of an ADS system's life-cycle, it includes development and design, procurement, deployment within a particular context, and the effects on downstream stakeholders. As we have seen from examples above what meaningful human control looks like and where it is best situated will depend on its purpose. Broadly speaking, for democratic engagement it will have at least a component in affected stakeholders, and for safety and reducing harm it better be situated in the deployment step, while for responsibility assignments, we might have more freedom how we distribute meaningful human control.

Meaningful human control is not a single solution for a single problem, but a tool for a variety of often unrelated problems that arise when using machine-learning AI and automated decision systems. The purpose of human control of AI should be explicitly stated and should drive institutional design. When the purpose of human control is clearly stated it can also provide guidance regarding the kinds of explainability that might be needed in a particular setting.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor EB declared a shared affiliation with the author at the time of review.

Publisher's note

References

Alan Turing Institute (2020). *Explaining Decisions Made with AI*. London: Alan Turing Institute. Available online at: https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf (accessed September 20, 2022).

Brennan-Marquez, K., and Henderson, S. (2019). Artificial intelligence and role-reversible judgment. *J. Crim. Law Criminol.* 109, 137. doi: 10.2139/ssrn.3224549

Brennan-Marquez, K., Levy, K., and Susser, D. (2019). Strange loops: apparent vs. actual human involvement in automated decision-making. *Berkley Technol. Law J.* 34, 745.

Brown, S., Davidovic, J., and Hasan, A. (2021). The algorithm audit: scoring the algorithms that score us. *Big Data Soc.* 8, 2053951720983865. doi: 10.1177/2053951720983865

Department of Defense (2016). *DoD Directive on AI Weapons*. Virginia: Department of Defense. Available online at: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf (accessed September 20, 2022).

Diamantis, M. (2021). Algorithms acting badly: a solution from corporate law. *Geo. Wash. L. Rev.* 89, 801. doi: 10.2139/ssrn.3545436

Ekelhof, M. (2018). *Autonomous Weapons: Operationalizing Meaningful Human Control*. Geneva: Humanitarian law and Policy, ICRC blog. Available online at: https://blogs.icrc.org/law-and-policy/2018/08/15/autonomous-weapons-operationalizing-meaningful-human-control/ (accessed September 20, 2022).

Elish, M. (2019). Moral crumple zones: cautionary tales in human-robot interaction. *Engag. Sci. Technol. Soc.* 19, 29. doi: 10.17351/ests2019.260

European Union Parliament (2016). *General Data Protection Regulation*. Strasbourg: European Union Parliament. Available online at: https://gdpr-info.eu/ (accessed September 20, 2022).

Felzmann, H., and Villaronga, E. F. (2019). Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* 6, 1–14. doi: 10.1177/2053951719860542

Green, B., and Kak, A. (2021). *The False Comfort of Human Oversight as an Antidote to Harm*. Washington, DC: Slate. Available online at: https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html (accessed September 20, 2022).

Horowitz, M. C., and Scharre, P. (2016). "A primer on meaningful human control in weapons systems," in *Paper for Center for a New American Security* (Center for a New American Security). Available online at: https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf (accessed September 20, 2022).

Larsson, S., and Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Rev.* 9, 1469. doi: 10.14763/2020.2.1469

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: mapping the debate. *Big Data Soc.* 3, 2053951716679679. doi: 10.1177/2053951716679679

Newman, J. (2021). *Explainability Won't Save AI*. Washington, DC: Brookings institute. Available online at: https://www.brookings.edu/techstream/explainability-wont-save-ai/ (accessed September 20, 2022).

Pasquale, F. (2021). *Inalienable Due Process in An Age of AI. Constitutional Challenges in the Age of AI*. Cambridge: Cambridge University Press, 42–56. doi: 10.1017/9781108914857.004

Purves, D., Jenkins, R., and Strawser, B. J. (2015). Autonomous machines, moral judgment and acting for the right reasons. *Ethic Theory Moral Pract.* 18, 851–872. doi: 10.1007/s10677-015-9563-y

Ribeiro, M. (2016). "Why should i trust you? Explaining the prediction of any classifier," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust*, 1135–1144. doi: 10.1145/2939672.2939778

Roff, H., and Moyes, R. (2016). "Meaningful human control, artificial intelligence and autonomous weapons," in *Briefing Paper for CCW* (UN Convention on Conventional Weapons). Available online at: https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf (accessed September 20, 2022).

Schwartz, E. (2018). *The (Im)possibility of Meaningful Human Control for Lethal Autonomous Weapons Systems*. Washington, DC: ICRC blog. Available online at: https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/ (accessed September 20, 2022).

Selbst, A. (2020). *Negligence and AI's Human Users*. 100 Boston Uni Law Review.

UNIDIR (2014). *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*. Geneva: UNIDIR. Available online at: https://unidir.org/publication/weaponization-increasingly-autonomous-technologies-considering-how-meaningful-human (accessed September 20, 2022).

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 17 February 2023
DOI 10.3389/fdata.2022.1019293

OPEN ACCESS

EDITED BY
George Lucas,
United States Naval Academy,
United States

REVIEWED BY
Craig Webster,
Ball State University, United States
Ricardo Crespo,
IAE Business School, Argentina

*CORRESPONDENCE
Sigurd N. Hovd
✉ sighov@prio.org

SPECIALTY SECTION
This article was submitted to Cybersecurity and Privacy, a section of the journal Frontiers in Big Data

RECEIVED 14 August 2022
ACCEPTED 15 November 2022
PUBLISHED 17 February 2023

CITATION
Hovd SN (2023) Tools of war and virtue–Institutional structures as a source of ethical deskilling. *Front. Big Data* 5:1019293. doi: 10.3389/fdata.2022.1019293

Tools of war and virtue–Institutional structures as a source of ethical deskilling

Sigurd N. Hovd[1,2]*

[1]Peace Research Institute Oslo (PRIO), Oslo, Norway, [2]University of Oslo, Oslo, Norway

Shannon Vallor has raised the possibility of ethical deskilling as a potential pitfall as AI technology is increasingly being developed for and implemented in military institutions. Bringing the sociological concept of deskilling into the field of virtue ethics, she has questioned if military operators will be able to possess the ethical wherewithal to act as responsible moral agents as they find themselves increasingly removed from the battlefield, their actions ever more mediated by artificial intelligence. The risk, as Vallor sees it, is that if combatants were removed, they would be deprived of the opportunity to develop moral skills crucial for acting as virtuous individuals. This article constitutes a critique of this conception of ethical deskilling and an attempt at a reappraisal of the concept. I argue first that her treatment of moral skills and virtue, as it pertains to professional military ethics, treating military virtue as a sui generis form of ethical cognition, is both normatively problematic as well as implausible from a moral psychological view. I subsequently present an alternative account of ethical deskilling, based on an analysis of military virtues, as a species of moral virtues essentially mediated by institutional and technological structures. According to this view, then, professional virtue is a form of extended cognition, and professional roles and institutional structures are parts of what makes these virtues the virtues that they are, i.e., constitutive parts of the virtues in question. Based on this analysis, I argue that the most likely source of ethical deskilling caused by technological change is not how technology, AI, or otherwise, makes individuals unable to develop appropriate moral–psychological traits but rather how it changes the institution's capacities to act.

KEYWORDS

virtue ethics, extended cognition, AI, automation, skill, deskilling, institution

Introduction

The prospect of autonomous artificially intelligent systems playing an increasingly prevalent part within the realm of warfare has been met by no small amount of concern from technologists, military ethicists, and the larger public alike. A central worry often expressed in this context has been the purported risk that the kind of automation these systems could facilitate and the risks of removing human ethical know-how from an area of paramount ethical concern. Former US Army Ranger Paul Scharre has provided, in his book *Army of None* (Scharre, 2018), a striking illustration of this conundrum from his own tours in Afghanistan in 2004. Here, Scharre describes working as a part of a

sniper team sent to the Afghanistan–Pakistan border to scout Taliban infiltration routes. While doing so, they are spotted by nearby villagers, and not long after a young girl, "of maybe five or six" heads out their way with a couple of goats on the trail (2018, p. 3). Under the poorly constructed cover of herding goats, the girl is to serve as a spotter for Taliban fighters. As Scharre notes, in this situation this young child was legally classified as a combatant, and was thus, technically, also a lawful target under the laws of war. But to treat her as such was to the team unthinkable. On discussing what would turn out to be a failed mission in its aftermath, no one brought such a course of action up as an eventuality: "We all knew it would have been wrong without needing to say it. War does force awful and difficult choices on soldiers, but this wasn't one of them" (2018, p. 4).

The question this story raises for Scharre is what happens with the nature of warfare when human agents are increasingly fighting their wars through and with AI systems and agents? One may ask this question through a variety of lenses. One lens would be to ask whether an AI system could ever be expected to make similar kinds of decisions and what such a system would look like.[1] Another lens is how the interaction with AI systems may come to change human beings' capacity for moral reasoning so that not even we can be expected to do so. That is, whether the implementation of and interaction with AI may come to change us in ethically problematic ways. From an ethical point of view, the latter of these lenses is arguably the most pressing. Not only does it address technology that is currently being implemented, but the question of whether an artificial moral agency is theoretically possible is a rather moot question if we are not capable of seeing the value in developing and implementing such systems. An influential concept through which this latter question has been raised is Shannon Vallor's concept of ethical deskilling (2015). As Vallor sees it, the introduction of AI technologies risks removing humans from the reality of war to a degree to which the skills constituting this moral wherewithal can no longer be developed.

While I think the notion of ethical deskilling to be a fruitful prism through which we can conceptualize some of the moral risks precipitated by this technological revolution, I will, in this article, argue we need to re-examine the notion of moral skill upon which this notion of deskilling is based. In so doing, I think there are good reasons to reject one central underlying assumption of Vallor, concerning both the nature and acquisition of these skills, as well as to rethink the most likely sources of ethical deskilling facing western militaries in the immediate future.

The article is divided into two main parts. In part one, I introduce Vallor's concept of moral skill and ethical deskilling and highlight some normatively problematic and moral psychologically implausible implications of a premise underlying this account. In part two, I go on to show how the concept can still be a highly relevant one if we take into consideration the institutionally mediated nature of professional military ethics.

Part 1: Vallor's concept of military virtue and moral deskilling

Moral skills: Some initial conceptual clarifications

Vallor's concept of ethical deskilling ties sociological literature on deskilling of workers within modern capitalism, originating in the work of Braverman, to neo-Aristotelian accounts of virtue represented by figures such as Hursthouse, McDowell, but most centrally, the study of Annas (McDowell, 1989; Annas, 1993; Hursthouse, 2002). Vallor's brief account of the concept of a moral skill, upon which this account of moral deskilling is supposed to rest, leaves, however, some ambiguity that will have to be addressed.

This ambiguity concerns how this notion of moral skill ties to the Aristotelian concept of moral virtue. Moral virtues are habitualized states of character disposing us to passion and action toward human flourishing. They are constituted as the mean between an excess and a deficiency (both vices), the way courage is the mean between cowardice and foolhardiness. The moral virtues move us to act by shaping our perception of the world (1109b). They allow us to recognize, in any given situation, what is good as good and what is bad as bad, good and bad here are understood in terms of human flourishing, and motivate us to realize the good (1113b). They are intrinsically tied to the intellectual virtue of practical wisdom, an intellectual virtue concerned with practical deliberation. Moral virtue provides us with the ends for which we act, and practical wisdom allows us to deliberate about the means through which they are enacted (1144a).

> Vallor's concept of a moral skill isolates the epistemic capacities tied to the moral virtues, both from their motivating capacity as well as their role in deliberation:
>
> Virtue must, therefore, be conceived as a habituated skill of discerning moral judgment joined with moral motivation and aim that guarantees the goodness of its use (Vallor, 2015, p. 110).

1 Asimov's laws of robotics is an attempt to answer this question from within the civilian sphere (Asimov, 1977). Stuart Russel's proposal for a design of a benevolent AI, and Bostrom and Yudkowsky's suggestion of designing such systems around decision trees, rather than neural networks, are yet others (Bostrom and Yudkowsky, 2011; Russel, 2019). The question of such designs for ethical AI could also be realized within the military space is an interesting ethical question, but as is about to become clear, it is not a central concern for this paper.

The way Vallor reads Aristotle, moral skills can be read as a stepping stone to proper virtue (Aristotle, 1984). They are, as she puts it:

> [A] sort of scaffold or stable grafting site upon which virtue can (but may or may not) take hold; for genuine virtue is something more than moral skill or know-how, it is a state in which that know-how is reliably put into action when called for, and is done with the appropriate moral concern for the good (Vallor, 2015, p. 110).

Vallor's claim that virtue is somehow founded on a set of moral skills rests largely on Aristotle's assertion that the performance of a virtuous act requires: (1) knowledge; (2) the right motivation; and (3) that it proceeds from an unchangeable character (Vallor, 2015, p. 110; Aristotle, 1105a). From this claim, Vallor seems to conclude that a virtuous act is the conjunction of a set of discrete moral psychological phenomena where moral skills are a foundational part. But this analytic claim from Aristotle, about what virtuous acts require does not in and of itself say anything about the moral psychological functioning of a virtuous person. It does not substantiate a claim about the existence of a set of moral skills upon which moral virtues are founded.

Some ambiguity attaches, then, to Vallor's concept of a moral skill. In specifying its content, she leans heavily on the Aristotelian and neo-Aristotelian moral psychology, but the concept also introduces structures that are hard to find ground for in this literature and may even contradict it.

Given Vallor's cursory treatment of the concept, my pragmatic solution to this problem is to treat the concept as referring to the epistemic capacities associated with the Aristotelian concept of moral virtues but to refrain from following Vallor in treating it as a separate and foundational moral psychological phenomenon. This will allow me to avoid the thornier exegetic questions raised above, while also making it easier to tie her account of deskilling to the larger virtue ethical tradition, a tradition that after all plays a central role in her larger thinking on ethics and technology. From a pragmatic point of view, I see this move as justified by the fact that Vallor's argument for viewing the argument that AI technology may be a source of deskilling, if successful, would hold equally well whether we treat the concept of a moral skill to pick out the Aristotelian concept of moral virtue or as a more fundamental moral epistemological phenomenon. For, while Aristotle makes a clear distinction between virtues (*aretes*) and skills (*technai*), he does claim that they are analogous in being taught through habituation. This is also the central feature of moral skills upon which her argument about the risks of deskilling rests.

Moral deskilling

The concept of deskilling originates from the sociological literature on work, in particular Braverman's *Labor and monopoly capital* (1974), and it refers to the process through which certain forms of skill and craftmanship can come to be made redundant and thus eliminated from work processes due to technological and managerially induced automation. Vallor's concept of moral deskilling refers to the elimination of moral skills from a given professional space through similar technological and managerial innovations. For moral skills to develop in an agent, they are dependent on factors such as exposure to models of the skill, basic motivation and cognitive and emotional resources, and a practical environment that allows sufficient opportunities for habituation. Within the military sphere, Vallor sees the processes of automation facilitated by AI technology as running the risk of hampering the habituation of moral skills among human military personnel by removing them from the realities of war.

> There exists a rich institutional tradition of education of virtue in most modern professional armies, especially within the officer or other leadership cores [...]. But as with all virtues they cannot be acquired in the classroom, or even in a simulator. Only in the actual practical context of war, where situations are neither stable nor well-defined and where success and failure have lifelong moral consequences, can words like "courage" and "discipline" be more than empty slogans or aspirational terms that cannot by themselves direct one to their achievement (Vallor, 2015, p. 114).

Vallor's worry, then, is that in a world where warfare increasingly is conducted by the means of technology that performs tasks that formerly only could be performed by humans, we will lose the moral wherewithal to use this technology in a virtuous fashion (Vallor, 2015, p. 115). Thus she questions, for example, whether human supervisors of a future army of lethal autonomous weapons could develop the right moral skills to provide any meaningful input to these weapon systems:

> A supervisor, in order to be legitimate authority, must have more experience and practical wisdom than the supervisee: in this scenario, what wisdom will future humans "on the loop" be able to offer the machines in this regard? (Vallor, 2015, p. 115)

While there very well may be something immediately intuitive about the idea that this kind of technological change can come to affect our capacity for moral cognition, I think there are good reasons to question one underlying premise of Vallor's particular argument for the potential for moral deskilling. Vallor's argument relies on the premise that military virtue is a *suis generis* form of virtue. Hence, military virtues like courage, discipline, honor, and respect have nothing in common with their non-military counterparts.

To see this, let us for a moment treat the military virtues as merely a subspecies of their respective general kinds so that

military and personal courage are two species of the same general kind of virtue, which is courage. Do we have the same reason to fear that human combatants, removed from the harsh realities of war are at risk of moral deskilling? I would argue, no. For if military virtue is not a *suis generis* kind of virtue, we must recognize that even a moderately virtuous person has a vast array of epistemic resources to come to recognize what is morally called upon them to do in a new situation. Their predicament is not to develop an entirely new set of character traits, but to learn how to apply their character to a new situation, i.e., to get an appropriate situational awareness of what is going on. For them to be able to do this, they might have to learn a new set of practical skills, but the fact that technological developments call on us to develop one set of practical skills rather than another does not mean that we are left incapable of recognizing right and wrong as a result.[2] This would only be the case if the necessary moral wherewithal required to make judgments about warfare had to be grounded in a very particular set of unmediated experiences of war. What this amounts to is treating military virtue as a *suis generis* form of moral cognition and virtue. I think there are good reasons to, on closer scrutiny, reject this notion both on normative and moral psychological grounds.

The notion that military virtue is a *suis generis* virtue is both normatively problematic and relies on questionable moral psychology. Assuming a broadly democratic outlook it is normatively problematic because it makes it difficult to see how one could justify any civilian control and oversight over the military if the normative standard on which the institution is to be judged are standard civilians who have no independent epistemological access to. How is, for example, the US Congress in any different position in relation to its military than the human supervisors are in relation to the army of lethal autonomous weapons mentioned by Vallor above? Furthermore, the moral psychological account of such a view of military virtue would seem to imply that it is simply highly implausible. If the appropriate moral skills necessary to act virtuously in a military setting can only be acquired within the context of war, then the number of such experiences afforded the average modern soldier would seem to highly underdetermine the presence of any habituated moral skill. Even adding the time that modern military personnel are given to reflect on the moral obligations attached to their professional roles, be it the regular soldier or even members of the officer core, it is hard to see how anything resembling bonified habituated virtues could be developed solely on this basis. The training would have to build upon preexisting moral skills. But if military virtue is a *suis generis* virtue, how could such pre-existing moral skills play any such role?

2 Even if military virtue only could be developed on the basis of moral skills of the kind we saw sketched out by Vallor above, and these skills could only be developed on the basis of a very specific set of experiences of war, in so far as these moral skills provide a foundation for military virtue, this would again imply that military virtue is a *suis generis* kind of virtue.

To illustrate the problem at hand, I wish to consider a case Vallor points to as a paradigmatic instance of military virtue: the case of Hugh Clowers Thompson Jr. and his actions during the My Lai Massacre. On 16 March 1968, US Army soldiers committed the mass murder of between 347 and 504 unarmed people in two hamlets, My Lay and My Khe, of the Son My village in Vietnam. While the incident was and is a blight on the reputation US army, it has also come to be known as a case of incredible moral heroism through the actions of Warrant Officer Hugh Clowers Thompson. The village Son My was suspected by the US army of being a Viet Cong stronghold. Thompson and his observation helicopter crew were given the task of assisting in a search and destroying the mission. The intelligence was wrong. Upon entering the village, the army was met with no resistance and no sign of the enemy. All the same, they went on to indiscriminately execute its population, men, women, and children. Thompson, upon realizing that a massacre was taking place did everything in his power to stop it, going so far as to land the helicopter between fleeing civilians and advancing land troops (Angers, 2014).

To Vallor, Thompson's actions are a perfect example of the kind of moral wherewithal human beings are capable of, even in as extreme circumstances as found in war. I wholeheartedly agree. But are the moral skills constituting this wherewithal grounded solely or even predominantly in experiences of war? A closer look at Thompson's life reveals many factors which may have had an impact on his capacity to embody such a heroic response, beyond his experience as a military officer. As Trent Angers has noted in his biography of Thompson, this is a person with an upbringing that in many ways prepared him for this moment (2014). Thompson was raised in a family environment that valued discipline and integrity. He had been a boy scout and had been actively involved in the Episcopal church. While being raised in Georgia, his family denounced the racism and discrimination taking place in the south. If we were to take all these factors into account–his strong ties to a religious community, the moral examples present in his life, and his engagement in organizations putting a strong value on service–is it impossible to imagine that Thompson would be able to manifest the kind of moral traits of character necessary to supervise military action mediated by lethal autonomous weapons? If we can imagine this, we need to rethink our account of ethical deskilling and the understanding of moral skills upon which it rests.

Part two: Moral skills and the social environment

On moral deskilling: Broadening the term

It is worth taking a closer look at the sociological origins from where Vallor appropriates the concept of deskilling. When Braverman, in *Labor and Monopoly Capital*, described the

introduction of scientific management theory (Taylorism) in the late nineteenth and early twentieth century as a process of deskilling, he is doing so with reference to labor as a social and political class (Braverman, 1974, p. 3, 294). The concept is closely tied to a Marxist understanding of capitalistic production and exploitation. Contrary to popular opinion, then as now, viewing western industrialized economies as tending toward a more and more skilled workforce, Braverman saw them as tending toward polarization. Capabilities earlier embodied in a class of craftsmen had been appropriated into the means of production. Management techniques, such as the division of production processes into ever smaller and more specialized labor tasks, lowered the skill required by any laborer, making each worker an ever more disposable part of the production process. The amount of knowledge put into the production process gets higher as industrialized societies are increasingly dependent on the skills of managers and engineers, but "the mass of workers gain nothing from the fact that the decline in their command of the labor process is more than compensated for by the increasing command on the part of managers and engineers" (Braverman, 1974, p. 294). To Braverman, then, his analysis of deskilling does not, first and foremost, refer to a loss of capabilities within a society or organization but to a change in the relation of power between labor and capital. To Braverman, deskilling refers to laborers' continuous loss of control over the production process *vis-à-vis* capital.

From a certain virtue ethical point of view, Braverman's account of deskilling makes for a compelling indictment of a capitalist economic system's capacity to provide a good foundation for human flourishing. In fact, it fits rather well with the anti-modern, neo-Aristotelian virtue ethics of Alisdair Macintyre (Macintyre, 2016, p. 91). If the notion of moral deskilling is to be applicable to the field of military professional ethics and professional ethics at large, however, some further translation work is needed. If we look at the types of organizations with a strong tradition for professional ethics, we are talking about institutions whose inner functioning cannot be captured in purely market economic terms. The purpose of these institutions, such as military, health and research institutions, judicial institutions, and so forth, is on a normative level to realize the common goods or values intrinsic to human flourishing. A military, for example, exists to safeguard the sovereignty of a state, maintaining its monopoly on violence within its territory by hindering interference from external threats. As such the legitimacy of its actions is tied up with the legitimacy of the state as a political entity. If a state's military conducts itself in a way that undermines the principles upon which the state's legitimacy rests, it is, at least on a normative level, undermining itself as an instrument of the state's political power.

Our theoretical interests in interrogating the possibility of ethical deskilling differ, therefore, in some crucial respects from the ones motivating the larger sociological literature on deskilling. Our worry that some crucial ethical wherewithal is being lost as ever-new facets of our active lives are being mediated by autonomous technology does not, first and foremost, pertain to the wellbeing of the professional, but the well-functioning of the institutions–to their ability to realize a common good. The reason we consider these skills inherently valuable has to do with the fact that the principles guiding the ethical wherewithal of these professionals are also constitutive of what it is that makes their institutions well-functioning. From this institutional perspective, it is therefore largely ethically unproblematic if ethical knowledge, which was prior embodied in the intuitive know-how of professionals, is now manifest in technological design and institutional structures. What is worrying is the risk of losing ethical knowledge on an institutional level.

To credibly substantiate this worry, we need an account of moral skills that: first, sheds light on the essential role played by the skill for the ethical functioning of the institution; and, second, how the introduction of a given technology may disrupt the functioning of this skill. In what follows, I will highlight a particular feature of the kind of ethical professional wherewithal theorized and cultivated within the field of professional military ethics that may help to provide the basis for such an account. I will argue that the kind of moral skills we want to see exercised by military professionals are essentially dependent on a larger institutional context, and this dependency makes them vulnerable to a particular kind of moral deskilling in the face of technological disruption.

Skills, deskilling, and the social environment

As skills are always embodied in concrete individual persons who possess them as capabilities to act, it is perhaps natural to think of the phenomenon of deskilling through an individualistic psychological frame. On such a view, we will think of deskilling as a process through which an individual either fails to gain or lose a capability through either a lack of practice or a change in other abilities underpinning the skill. I may never become a good trumpetist because of a lack of disciplined training. I may once have been a decent trumpetist but let the skill atrophy. Or, alternatively, I may no longer be a great trumpetist due to an injury to my lung. But skills, as capabilities to act, are arguably almost always environmentally embedded phenomena. In describing what I am capable of doing, I am usually also making some reference, however implicit, to an environment facilitating and sometimes also partly constituting these capabilities. In analyzing the activity of tying one's shoelaces, one must make some kind of reference to the laces being tied, and thus, this environmental feature can be said to be a constitutive part of the skill of

shoe tying. The environmental features that can facilitate or constitute a skill are not restricted to the physical environment. Professionals such as stock traders, auctioneers, and rhetoricians possess skills facilitated by the social environment in which they were developed, skills that may rely on features of this social environment to a degree to which they can be said to partly constitute what it means to perform them all.

If we take these forms of environmental dependency into account, new possible forms of deskilling also come to the fore. As these skills usually are dependent also on environmental variables, changing these variables can change the individual's capabilities. At a height of 100 m above sea level, I may be a decent trumpeter, but less so in the low air pressure of high altitudes. I may be a skilled rhetorician under certain social circumstances but not under others. In this way, changes in the physical and social environment in which an agent is embedded may either enhance or diminish the skillfulness with which they are able to face a given situation.

The skills of professional athletes are an interesting example in this regard. To be a skilled professional basketball player, cross-country skier, or runner involves interacting with an environment that is socially and technologically mediated, where what it means to be a skilled athlete, in any of these sports, is in part determined by its social and technological features. In defining standards of excellence within the game of basketball, for example, we must make some reference, either implicitly or explicitly to the rules of the game. Throughout the history of the game, these rules have changed. Many of these changes, instituted by leagues such as the NBA or the NCAA, have been made in order to either enhance or diminish certain star players playing within the league at the time. To mention a few, in 1947, the NBA banned zone defense in order to enhance the impact of dominant players like Neil Johnston, Dolph Schayes, and Bob Petit (Warond, 2017); in 1951 and 1964, the league expanded the lane, first from 6 to 12 feet and then from 12 to 16, encouraging guard and wing play and curbing the power of taller star players such as George Mikan and Wilt Chamberlain. From 1967 to 1976, the NCAA banned dunking, responding allegedly to the strength of Lew Alcindor (Caponi, 1991, p. 4); and in 2001, the NBA reinstated the zone defense to offset the pure physical dominance displayed by Shaquille O'Neil (Warond, 2017).

Such changes in the conventional norms governing the game may, from the perspective of the individual player, be described as an instance of deskilling. A player may have developed skills throughout their career, the significance of which may be greatly diminished as a result of the changing rules. Physical attributes which at one point gave them a crucial advantage may no longer be as significant, and new skills and physical attributes may come to the fore as essential to the game. It may also be described as an instance of deskilling from the perspective of the team. Basketball is a team sport, and the significance of any particular skill in making a good basketball player is determined by its role in the collective effort of winning the game. In light of a significant change in the rules of the game, roles and tactics may become obsolete, and a team, once pre-eminent, may find their style of play an ill-fit in this new environment.

Within sports, as in so many other instances of expertise, this kind of deskilling through environmental change is often precipitated by technological development. In cross-country skiing, for example, changes in ski design and vaxing technology have completely changed the physical profile of elite professional skiers, now increasingly favoring athletes with a higher muscle mass. If a professional skier from the 1970's was transported to the world championship, anno 2022, they would consequently suffer from deskilling purely because of changes in the technological environment they are interacting with. For the Nordic countries, who for cultural and historic reasons have invested heavily both in skiing and vaxing technology as well as empirical research into professional development, these changes in the technological environment have given them a comparative advantage, leaving other once strong skiing nations in the dust.

Military ethics and institutional environments

If we are to look for potential threats of moral deskilling within the military as a result of technological disruption, I would propose that the most pressing threats are analogous to the ones presented above; not precipitated by changes in the moral character of military personnel seen in isolation, but by changes in their environment, in particular, the institutionally mediated environment in which military personnel are embedded. To defend this view, I will in this section provide a brief account of the role of institutional mediation in military ethics, and in the next section provide an example of how this institutional context can be a source of deskilling.

This notion of an institutionally mediated environment calls for some further conceptual clarification. By an institution, I mean to refer to an organization constituted by an embodied structure of differentiated roles (Miller, 2010). These roles are defined in terms of a set of interdependent tasks with reference to a shared overarching end, a set of formal rules regulating suitable behavior, as well as a set of informal cultural norms. Institutional behavior, like all human behavior, is also supported by a set of material and technological conditions. In talking about an institutionally mediated environment, I mean to refer to the space of action of agents embodying a given role within an institution, for example, how the institution through rules, cultural norms, and material and technological resources both constrains and extends the agent's own capabilities to act. Institutions are tools for collective action, they enable individuals acting through them to perform feats far beyond the power of any singular person. With this power comes moral obligations and social expectations which constrain how this role may be wielded in a morally and socially legitimate way.

The account of deskilling I wish to put forward assumes that moral skills associated with the military profession–skills usually described as military or martial virtues–are species of their more general kinds rather than a sui generis form of moral traits. Military courage is simply a specific form of courage, military loyalty is a specific form of loyalty, and military prudence is a specific form of prudence. Additionally, following a loosely Thomistic account of military virtue, I would argue that what characterizes these virtues as virtues of a specific kind is that they are exercised in the pursual of a common good–a precondition of human flourishing–that cannot be realized by any one private individual, i.e., securing the safety and autonomy of a polity against external threats.

Similar to Aristotle, St. Thomas saw the moral virtues, i.e., dispositions toward the good, as virtues proper, only in so far as they were guided by the intellectual virtue of practical wisdom, or prudence–excellence in deliberation practical deliberation for the sake of the good (1140a; ST I-II q47 a 2). One might be disposed to act in ways that happen to be courageous or temperate by means of a good upbringing or natural dispositions. However, unless oneself am capable of deliberating about the ends for which my nature and upbringing disposes of them, one is not truly courageous or temperate (1144b14-18; ST I-II q 65 a 1). Through virtue, I recognize and am inclined toward the good and through practical wisdom, I deliberate about means to realize it, specifying its content in this particular case. But the good toward which virtue directs us and our practical wisdom deliberates about in military affairs is one that can be realized solely by our own actions. It must be pursued by the polity collectively. Military prudence, then, because it is ultimately concerned with the safety and autonomy of the polity, a good that no single member of the polity has the power to realize by themselves, may only be exercised if an agent's practical deliberation is integrated into a larger collective project. The institution of the military is a tool enabling the integration of individual actions into such a larger project. Additionally, military prudence is a virtue concerned with practical deliberation regarding collective action, as mediated by this institution, for sake of the common good.

As argued by Reichberg (2017), St. Thomas was thus also well aware of the essential role played by institutional mediation in the exercise of prudence within the military sphere. We see this reflected both in his understanding of political and military authority as being legitimized by the need for a coordinated endeavor to be guided by a unified will (ST II-II q 40a1), as well as in his treatment of the kinds of prudence required by humans engaged in acts of war.

St. Thomas' conception of military prudence is closely tied to his conception of prudent governance more broadly, as well as his understanding of the nature of legitimate political authority. Working from the premise that there are common goods that only can be realized through coordinated collective action, he consequently emphasizes a distinction between the prudence required of a ruler (*prudentia regnitiva*) and the prudence required of the ruled (*prudentia politica*), as well as recognizing a distinct form of *prudentia regnitiva* pertaining to military affairs (STII-II q 50 a4). For rational agents engaged in such collective endeavors, there exist, according to Aquinas, different standards of deliberative excellence, determined by the agent's role in the collective endeavor: a virtue of command and a virtue of obedience (Aquinas, 1981). But as Reichberg notes, obedience, as a virtue, is never blind. By treating both a commander's giving of orders and a subordinate's implementation of them as instances of deliberation and action falling under the confines of the intellectual virtue of prudence, they are both realized only insofar as they are guided by the common good:

> [T]o be humanely exercised civic obedience requires a special part mode of deliberation; in this respect, purely personal prudence is an incomplete guide, for, in addition to reflecting on the implications for my private good, I must weigh the concordance of the command with the common good of the polity of which I am a member (Reichberg, 2017, p. 140).

Both the commander's and the soldiers' actions, then, are constitutive parts of a unified collective act that must be guided by the common good, as far as they are to be considered moral. For their actions to be guided by the common good means for both to be correctly attuned to the institution of which they are a part, an institution that mediates their actions through a set of differentiated roles.

We commonly do not think of institutional structures as being an essential part of moral reasoning. When, for example, Kant, in the Critique of Practical Reason (Kant, 1997, p. 118/5, 148), speaks of the moral law within, he is describing a fairly widely held picture of moral conscience and cognition. That is, as a process taking place in the deepest privacy of our own minds, our humanity, pure and simple, stripped of anything accidental, like sentiment or convention. From within such an internalist picture, it is perhaps natural to say that the best we may ask of our institutions is to not stand in the way of our conscience. Indeed, Hannah Arendt has shown us how institutional structures can come to be deeply detrimental to our capacity to recognize the call of morality, describing in her study of Adolf Eichmann, how legal and bureaucratic structures can facilitate acts of evil whose terribleness is matched only by their banality (Arendt, 1963, p. 72, 118). Yet, what St. Thomas's accounts of the ethics of war so perfectly illustrates is that our moral conscience does not always appeal to our humanity simpliciter. Sometimes its appeal is an appeal to us in virtue of our profession, as a nurse, a journalist, a judge, or a soldier.[3] These roles are partly defined through the institutional

3 That social roles might be ethically informative is in one sense clearly recognized. Role ethicists such as Baril, Garcia, and Scheffler have all

structures into which they are integrated. Structures play a constitutive role in determining what ethical action means within a professional space because they determine what it means to embody a given professional role–invested by a society with unique powers and ethical obligations.

A useful prism to understand the role played here by these institutional structures is the concept of extended cognition. In their seminal article from 1998, *The Extended Mind,* Andy Clark and David Chalmers suggested that features of our external environment may not only facilitate processes such as calculation, memory, belief formation, orientation, and reasoning but may as well come to serve constitutive parts of these processes themselves. Cognitive tools, such as notebooks, cellphones, and abacuses, can thus be said to extend our mind beyond our body and into the environment itself (Clark and Chalmers, 1998). As was already argued then, there is no principled reason to think that parts of our social environment cannot play a similar function:

> Could my mental states be partly constituted by the states of other thinkers? We see no reason why not, in principle. In an unusually interdependent couple, it is entirely possible that one partner's beliefs will play the same sort of role for the other as the notebook plays [. . .] (Clark and Chalmers, 1998, p. 17).

More recently Shaun Gallagher and Anthony Crisafi have pointed to institutional structures, the ones constituting the institution of the law, as an instance of socially extended cognition in some ways even more radical than what Clark and Chalmers immediately had in mind:

> There may be external resources that can carry out cognitive processes that in principle may not be possible to do in our head, and that we would have a hard time conceptualizing as something we could even refer to the phrase "if it were done in the head" (Gallagher and Crisafi, 2009, p. 47)
>
> The law, which may be the product of previous generations, but is currently organized in legal institutions, operates like a mechanism that helps to accomplish our thought (Gallagher and Crisafi, 2009, p. 48).

The work of Clark, Chalmers, Gallagher, and Crisafi may suggest a promising perspective to think of the nature of the so-called martial virtues, as concerned with the attunement of our moral character to a unique cognitive tool: the military institution. Of course, much work remains in exploring the ultimate viability of applying this conceptual framework to the topic of professional military ethics–work that would take us too far afield from the aims of this present paper.[4] What immediately can be achieved by taking up such a perspective, however, is to put the collective technologically mediated nature of military action in focus. Thus, if nothing else, it can function as a corrective to a moral psychological point of view still dominating most contemporary moral philosophy. Seeing professional military ethics through the prism of extended cognition invites us to the words of Clark, "to cease to unreflectively privilege the inner, the biological, and the neural" (Clark, 2011, p. 218), and as I hope to show in the next section, it can shed light on how our capacities for moral reasoning can be undermined by technological change.

Military ethics and moral deskilling

Already in the short to medium term, the technologies falling under the rubric of AI holds the promise of issuing an era of unprecedented automation. So too in the military sphere, different levels of autonomy have already been introduced in a vast array of weapon systems and continue to do so in the future, opening new tactical and strategic advantages whose ramifications we are only beginning to understand. With the introduction of this technology comes also new challenges. One challenge I here wish to bring to the fore pertains to the scalable nature of autonomous weapon systems. Stuart Russel, who otherwise has expressed skepticism about the moral panic about lethal autonomous weapons, has nevertheless expressed some worry about just this feature:

> [A] process is scalable if you can do a million times more of it with buying a million times more hardware. Thus Google handles a roughly five billion search requests per day by having not million of employees but million of computers. With autonomous weapons you can do a million times more killing by buying a million times more weapons precisely because the weapons are autonomous. Unlike remotely piloted drones or AK-47s, they don't need individual human supervision to do their work (Russel, 2019, p. 112).

highlighted ways in which social roles can come to guide moral behavior (Garcia, 1986; Scheffler, 1997; Baril, 2016). Recently, Joseph Chapa has also argued, I think concvincingly, that this is form of ethics may shed an interesting light on the nature of the martial virtues, and the two often seemingly conflicting normative demands placed on the soldier, that practical efficiency and moral integrity (Chapa, 2018).

4 Treating a mode of moral cognition as dependent on features in the environment that is historically contingent, brings up meta-ethical challenges that such a view will have to address. Perhaps the most pressing question one will have to answer is how it is that we expect such a cognitive process to generate judgements with the kind of universal force we commonly associate with moral judgements.

If we were to borrow some quasi-Marxian economic terms, we could say, then, that AI holds within it the potential of transforming vast amounts of the human labor involved in the process of military action into pure "means of destruction" through the process of automation. This obviously constitutes a potentially massive expansion of the destructive potential of modern militaries; first of all, because labor is a much scarcer resource than capital; second, because the loss of one's own soldiers, is one of the biggest political restrictions for a polity to revert to force. I would argue that it is also a potential source of ethical deskilling.

The scalable nature of autonomous weapons means their introduction into a military will entail a massive expansion of the potential scale of the institution's actions. Seen in isolation, the actions it affords the institution might not seem qualitatively different from those already in its arsenal. But increasing the scale on which an institution may act can come to change the nature of these actions themselves. It is not given that individual humans operating within the institution will possess sufficient conceptual frameworks to apprehend the ethical implications of this increase in scale.[5] The introduction of AI technology into military institutions may thus cause ethical deskilling among military professionals by changing the institution within which they act. They might possess the exact same personality traits, physical capabilities, and self-understanding as they did prior to the introduction of this technology. But just like the way a change in the rules of basketball can come to change what it means to be a good basketball player, a technological change might change the traits, capabilities, and self-understanding needed to attune oneself to the institution of the military in an ethical fashion.

Arguably, a recent example of this kind of technologically and institutionally grounded ethical deskilling is found in the use of unmanned aerial vehicles (UAVs or drones) as a part of the US government's war on terror. While Russel is right in highlighting the unique threat represented by the scalability of autonomous weapon systems, when seen from a more sociological perspective, what drone warfare represents to modern militaries bears strong analogies with the benefits represented by the scalability of autonomous weapons. Drones increase a state's air power by lowering demands put on the personnel operating the aircraft, both in terms of the training needed for a human to be able to man it and the risks subsequently placed on this asset (drone pilots are not in immediate risk of being shot down). Drone warfare has thus radically lowered the cost of both applying and projecting air power, financially, manpower-wise, and politically. As was born out in the US war on terror, this lowering of the cost of action enables an increase in scale.

Actions once increased in scale can come to undergo metamorphoses, not always predictable to those putting them into being. So we arguably saw this reflected in the Obama administration's legal and moral justification for the use of UAVs to take out members of Taliban and Al-Qaeda leadership. When the US State department's Legal Advisor Harold Koh first presented this justification, he argued that there is nothing about this technology, *per se*, that would make it contrary to the laws of war. Nor is there anything unprecedented about the way this technology had been used by the US armed forces, as the US has used airstrikes to take out enemy leadership going back to the second world war. In fact, they argued, the targeted nature of such strikes enables the US to wage this war in a way that is both limited and proportional. One obvious objection here is that to consider the technology *per se* is to consider an abstraction. It does not allow us to come to terms with the question of proportionality because it does not allow us to perceive scale. It does not allow us to see the cost for a civilian populace living under the watchful gaze of tools of death and destruction, loitering in the sky above them. It does not invite the question of whether in the name of a war on terror one would be justified in placing these people under a *de facto* reign of terror.

By an increase in the scale of the action in question, then, the moral obligations associated with the action can change as well. And there is no guarantee that any one operator supervising these weapon systems has the sufficient overview to catch such changes, as well as the normative frameworks to problematize them. The same goes for the decision apparatus that exists within the institution to regulate and supervise their actions and civilian institutions in charge of providing external oversight.

Consequently, then, when worrying whether UAVs and LAWs may precipitate a loss of capabilities for moral reasoning, it may be natural to formulate one's worries as a question of whether drone operators or artificially intelligent autonomous systems possess sufficiently similar qualities to human beings now performing similar actions. But as war is an intrinsically collective endeavor, professional military ethics cannot be concerned solely with individual actions, abstracted from the larger collective endeavors of which it is a part. The question of military deskilling must therefore be treated as an institutional, rather than a purely individual concern. The question we must ask ourselves is whether the decision structures governing these intrinsically collective actions will possess the right kind of knowledge, an incentive structures to see and respond to a changing moral landscape. Seen from this perspective, the more likely candidates for ethical deskilling may not first and foremost be found among the average operators, but within military leadership. It is here, after all, that responsibility for these collective actions as a whole resides.

5 Of course, this is only one of many ways in which the introduction of AI technology may come to change the institutional environment through which we act.

Conclusion

Vallor's notion of ethical deskilling is a powerful one, capturing a deep-seated ambiguity about technological change and autonomous systems within the military sphere. I have in this article looked closer at the seams of the conception of a military virtue underpinning Vallor's application of this concept: whether military virtue is a *suis generis* form of virtue. I have argued that we have reasons to reject it both on normative and moral psychological grounds. From a normative perspective, this view of military virtue seems to imply a kind of moral exceptionalism on behalf of military institutions within their domain from where it would be hard to justify any kind of civilian oversight or control. From a moral psychological point of view, it is simply implausible, as the experience of combat afforded the average soldier woefully underdetermines the emergence of any independent moral skill. Military virtue must in some way be grounded in virtue simpliciter if it is reasonable for us to expect it to develop at all.

I have subsequently offered an alternative account of ethical deskilling on the basis of a view of military virtue and professional military ethics as a particular form of extended cognition. On this account, what makes military virtue a specific kind of virtue is its mediation through the military institution. Military virtue is virtue exercised on behalf of a common good that can only be realized through coordinated collective action. The military as an institution exists to coordinate this collective effort and functions as a cognitive tool through which individuals can partake in it. I have proposed that the perhaps most plausible sources of ethical deskilling are grounded in just this institutional-embedded nature of military virtue. Technology changes the capabilities of institutions in ways that may neither be entirely predictable to those introducing it nor to people set to use it. Radical technological change such as the one promised with the advent of ever new and more complicated AI technology can, thus, become a source of ethics by disorienting the institution to the ethical implications of their collective effort.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

The paper was written solely by SH.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Angers, T. (2014). *The Forgotten Hero of My Lai: The Hugh Thompson Story (Revised edition)*. Lafayette, Louisiana. Acadian House Publishing.

Annas, J. (1993). *The Morality of Happiness*. Oxford: Oxford University Press.

Aquinas, T. (1981). *Summa Theologia*. Vol. 4. Translated by Fathers of the English Dominican. Province, Westminster, MD: Christian Classics.

Arendt,. H. (1963). *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: The Viking Press.

Aristotle (1984). *Nicomachean Ethics*. Ross, W. D. (trans.), revised by Urmson, J. O., in *The Complete Works of Aristotle*, The Revised Oxford Translation. eds Barnes, J. Princeton: Princeton University Press.

Asimov, I. (1977). *I, Robot*. New York: Fawsett Grest.

Baril, A. (2016). The ethical importance of roles. *J. Value Inq.* 50, 721—734. doi: 10.1007/s10790-016-9575-4

Bostrom, N., and Yudkowsky, E. (2011). *Cambridge Handbook of Artificial Intelligence*. eds Ramsey, W., and Frankish, K. Cambridge: Cambridge University Press.

Braverman, H. (1974). *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*. 25th Anniversary Edition. New York: Monthly Review Press.

Caponi, G. D. (1991). *Signifyin(G), Sanctifyin', and Slam Dunking: A Reader in African American Expressive Culture*. Boston: University of Massachusetts Press.

Chapa, J. O. (2018). A roe morality for soldiers? *Soc. Theory Pract.* 44, 179–198. doi: 10.5840/soctheorpract20183 1533

Clark, A. (2011). *Supersizing the Mind: Embodiment Action and Cognitive Extension*. New York: Oxford University Press.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis.* 58, 7–19. doi: 10.1093/analys/58.1.7

Gallagher, S., and Crisafi, A. (2009). Mental institutions. *Topoi* 28, 45–51. doi: 10.1007/s11245-008-9045-0

Garcia, J. L. A. (1986). Morrally ought rethought. *J. Value Inq.* 20, 83–94. doi: 10.1007/BF00144536

Hursthouse, R. (2002). *On Virtue Ethics*. Oxford: Oxford University Press.

Kant,. I. (1997). *Critique of Practical Reason*. eds Gregor, M. J. Cambridge: Cambridge University Press.

Macintyre, A. (2016). *Ethics in the Conflicts of Modernity: An Essay on Desire, Practical Reasoning, and Narrative*. Cambridge: Cambridge University Press.

McDowell, J. (1989). *Mind, Reality and Value*. Cambridge: Harvard University Press.

Miller, S. (2010). *The Moral Foundations of Social institutions: A Philosophical Study*. New York: Cambridge University Press.

Reichberg, G. (2017). *Thomas Aquinas on War and Peace*. New York: Cambridge University Press.

Russel, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking Press.

Scharre, P. (2018). *Army of None*. New York: W. W. Norton and Company.

Scheffler, S. (1997). Relationships and responsibilities. *Philos. Public* Aff. 26, 189–209. doi: 10.1111/j.1088-4963.1997.tb00053.x

Vallor, S. (2015). Moral deskilling and upskilling in a machine age: reflections on the ambiguous future of character. *Philos. Technol.* 28, 107–124. doi: 10.1007/s13347-014-0156-9

Warond, A. (2017). 5 Players Who Forced The NBA To Change Their Rules. *Fadeaway World*. Last modified June 2 2017. Available online at: https://fadeawayworld.com/more-than-points/5-players-who-forced-the-nba-to-change-their-rules (accessed July 30, 2022).

frontiers | Frontiers in Big Data

TYPE Conceptual Analysis
PUBLISHED 12 May 2023
DOI 10.3389/fdata.2023.1020107

OPEN ACCESS

EDITED BY
Henrik Syse,
Peace Research Institute Oslo (PRIO), Norway

REVIEWED BY
Sarah Michele Rajtmajer,
The Pennsylvania State University (PSU), United States
Dalton Lunga,
Oak Ridge National Laboratory (DOE), United States

*CORRESPONDENCE
Mitt Regan
✉ regan@georgetown.edu

†These authors have contributed equally to this work

RECEIVED 24 August 2022
ACCEPTED 25 April 2023
PUBLISHED 12 May 2023

CITATION
Regan M and Davidovic J (2023) Just preparation for war and AI-enabled weapons. *Front. Big Data* 6:1020107. doi: 10.3389/fdata.2023.1020107

Just preparation for war and AI-enabled weapons

Mitt Regan[1*†] and Jovana Davidovic[2†]

[1]Georgetown Law, Georgetown College, Georgetown University, Washington, DC, United States, [2]Philosophy Department, The University of Iowa, Iowa City, IA, United States

This paper maintains that the just war tradition provides a useful framework for analyzing ethical issues related to the development of weapons that incorporate artificial intelligence (AI), or "AI-enabled weapons." While development of any weapon carries the risk of violations of jus ad bellum and jus in bello, AI-enabled weapons can pose distinctive risks of these violations. The article argues that developing AI-enabled weapons in accordance with jus ante bellum principles of just preparation for war can help minimize the risk of these violations. These principles impose two obligations. The first is that before deploying an AI-enabled weapon a state must rigorously test its safety and reliability, and conduct review of its ability to comply with international law. Second, a state must develop AI-enabled weapons in ways that minimize the likelihood that a security dilemma will arise, in which other states feel threatened by this development and hasten to deploy such weapons without sufficient testing and review. Ethical development of weapons that incorporate AI therefore requires that a state focus not only on its own activity, but on how that activity is perceived by other states.

KEYWORDS

automated, weapons, testing, security, war

Introduction

Emerging attention to *jus ante bellum* as an element of the just war tradition reflects attention to "just preparation for war." As Ned Dobos frames the issue, "When (if ever) and why (if at all) is it morally permissible to create and maintain the potential to wage war?" (Dobos, 2020, p. 2). We agree with Cecile Fabre that maintaining a standing army that is prepared to wage war if need be is morally justified because it enables a state to protect persons from violent infringements of their fundamental rights (Fabre, 2021). We argue, however that *jus ante bellum* still requires a state to morally justify the particular ways in which it engages in such preparation. Harry van der Linden suggests that this requires that a state prepare for war in ways that minimize the risk of unjust resort to force—violations of *jus ad bellum*—and unjust use of force during war—violations of *jus in bello* (van der Linden, 2010, p. 7).

This essay examines what *jus ante bellum* requires of states regarding the development and deployment of weapons enabled by artificial intelligence (AI). We define these as weapons that utilize artificial intelligence and machine-learning models in the targeting process, which may include tasks such as object recognition, target identification, or decision-support. We focus on the targeting process, and define AI-enabled weapon systems as those that use AI in that process, because the human-machine interactions in the targeting stages have the most consequential effects on war and on the ways in which norms of war may be violated (Ekelhof, 2018). To clarify, the targeting process consists of several steps at which humans and machines may interact in complex ways, with machines augmenting

rather than displacing human judgment. But even when a human is the ultimate decision-maker at the last step, these interactions can shape their understanding of the situation they confront in powerful ways, which in turn influences their decision as to whether to fire.[1]

We believe that, in light of increasing attention by several states to the potential for incorporating AI into weapon systems, a state is justified in investing in developing such systems in order to protect its population [see Boulanin and Verbruggen (2017) for a discussion of the current state of such efforts]. We argue, however, that *jus ante bellum* requires that before deploying these weapons a state must engage in a rigorous testing, evaluation, verification, and validation (TEVV) process, which we describe below. It must also carefully consider the appropriate delegation of tasks between machines and humans.[2] Finally, it must engage in development of these weapons in ways that do not trigger a security dilemma that leads other states to deploy AI-enabled systems without engaging in these processes.

These requirements reflect concern that premature deployment of AI-enabled weapon systems, and the deployment of systems with an inappropriate delegation of authority between machines and humans, increase the risk of violations of *jus ad bellum* and *jus in bello*. The next section elaborates on these risks.

Risks of AI-enabled weapons

Aside from the risks that arise in any complex tightly coupled system, AI-enabled weapons have at least two features that can pose distinctive risks. First, systems at this point tend to be brittle, in the sense that they are not able to function effectively outside the specific set of circumstances for which they are programmed. It can be challenging for operators to identify when this happens, and to predict the consequences. Second, a system may not be able to provide an explanation of its analysis and recommendations in terms that are comprehensible to a human operator. This opacity can make it difficult for humans to exercise effective judgment about potential courses of action.[3]

These features of AI-enabled weapon systems could increase the risk of violations of *jus ad bellum* and *jus in bello*. With respect to *ad bellum*, states could field systems that are less flexible than conventional weapons and lack sensitive contextual awareness of likely human intentions. "This brittleness of machine decision-making may particularly be challenging in pre-conflict crisis situations, where tensions among nations run high," and contextual human judgment can be crucial in lessening the risk of escalation (Horowitz and Scharre, 2021). Furthermore, even if a system performs as intended, adversaries may not know whether its behavior reflects human intention. This ambiguity may lead to escalation of conflict if states assume that they must ascribe hostile intention to an adversary in order to protect themselves.

With respect to *in bello* violations, delegation of some tasks to machines could mean that "minor tactical missteps or accidents that are part and parcel of military operations in the chaos and fog of war, including fratricide, civilian casualties, and poor military judgment, could spiral out of control and reach catastrophic proportions before humans have time to intervene" (Horowitz and Scharre, 2021). This risk would be exacerbated by the interaction between and among competing AI-enabled systems, which could result a cycle of attacks and counterattacks at a speed that humans could not control.

These risks underscore the crucial importance of rigorous pre-deployment review of AI-enabled weapons. States ordinarily would have incentives to engage in such review to ensure that they can exercise effective control of these weapons. Their willingness to do so could be lessened, however, by what is called the security dilemma. This occurs when states perceive that other states' military investment make them less secure, a perception that may be especially likely because of the perceived decisive advantage that AI-enabled weapons can provide. *Jus ante bellum* therefore requires not only that states not deploy AI-enabled systems without rigorous TEVV, but that they engage in development of these systems in ways that minimize the risk of a security dilemma. The next section discusses what states can do to conduct rigorous TEVV, while the following section discusses how they might take steps to avoid triggering a security dilemma.

Testing, evaluation, verification, and validation

Deployment of AI-enabled weapons that have not been rigorously tested for safety and reliability would increase the risks of unjust resort to war and harm to innocent persons. To avoid these risks, deployment should be preceded by rigorous engagement in a process known as testing, evaluation, verification, and validation (TEVV). This process, drawn from systems engineering, is designed to assess the future performance of new technology and the risks that it may pose. While TEVV is the most common description of the steps in this process, terminology can vary, and the steps themselves are not strictly separate.

In the defense setting, the Department of Defense Instruction on Test and Evaluation (T&E) says, "The fundamental purpose of T&E is to enable the DoD to acquire systems that support the warfighter in accomplishing their mission" [UD Department of Defense, 2021, §3.1(a)]. Verification seeks to ensure that the technology meets the specifications that a prospective user has provided, while validation assesses whether those specifications will meet the goals of the user (Hand and Khan, 2020).

1 Furthermore, while, for example, an autonomous driving tank might be AI-enabled in some sense it raises very different issues than those "AI weapons" that use AI for primarily for war-fighting purposes.

2 Various sources have the "VV" as "validation and verification" (Flournoy et al., 2020), while NSCAI (2021) and DoD AI strategy documents have it as "verification and validation." Here we use VV to mean verification and validation, partly because we see sources such as NSCAI as authoritative in the U.S. context, and partly because validation is the last step in the process in which machine-learning models are built and tested (We thank Joe Chapa for this clarification).

3 Careful TEVV can uncover such explainability limitations and that can in turn inform potential remedies to inscrutability; remedies that might include re-training of operators, changing the user interface, augmenting the algorithm with XAI tools, or in some cases when such lack of explainability significantly negatively affects calibrated trust in operators, abandoning the algorithm.

The TEVV process thus seeks to provide assurance that technology will work as expected, which generates what Roff and Danks call predictability-based trust (Roff and Danks, 2018). Because a weapon can cause significant harm, however, TEVV of weapon systems also must provide what Roff and Danks call values-base trust: confidence that a weapon will operate in a way that is consistent with relevant ethical principles.

As Roff and Danks observe, the challenge is that the paradigm of values-based trust is interpersonal relationships, in which trust reflects confidence that another person will act ethically in unpredictable future situations because we know the values and beliefs that guide them (Roff and Danks, 2018, p. 7). Developing such trust in a machine is much more difficult. Yet the more advanced an AI-enabled weapon system, the more crucial the need to trust that the outputs of its automated components are consistent with ethical principles.

TEVV thus must seek to foster the right kind of calibrated trust in commanders who decide to deploy the weapon system and operators who use it. Trust is *calibrated* when the degree of reliance is appropriate to the system's predictable performance in a particular context (Pinelis, 2021). Trust is of the *right kind* when it is grounded not only in predictability but in confidence that a system will operate in conformity with appropriate ethical values (Roff and Danks, 2018). This can be achieved partly by embedding ethical considerations into the TEVV process and assuring operators and commanders that the legal review and TEVV process not only assures predictable performance, but predictable performance in accord with, for example, *jus in bello* principles.

AI-enabled weapons present distinctive challenges for the TEVV process because of their complexity, opacity, and brittleness. While we cannot discuss all these challenges here, we discuss especially significant ones below, and suggest how TEVV should respond to them in order to satisfy *jus ante bellum*.

Challenges

Generalizing and extrapolating from test results is especially difficult for many AI-enabled weapon systems because of the exceptional difficulty in anticipating all the conditions under which these weapons will operate. It is true that conventional weapons present a similar obstacle to some extent, since we can test only a fraction of the settings in which a weapon may operate. AI-enabled weapons, however, perform extremely complex tasks, they do so in radically unpredictable environments, and they provide "non-deterministic, dynamic responses to those environments" (Wojton et al., 2021, p. 4). Their likely failures also will be harder to predict and understand than those of conventional weapons. All this makes the range of potential scenarios to test immense, if not infinite.

Compared with conventional weapons, we therefore will be able to generalize with less confidence about performance across varied environments, and less easily identify settings to which the use of a weapon should be confined (Pinelis, 2021). In addition, it may be necessary to move away from insistence on complete risk avoidance and precise risk quantification toward acceptance of some risk of failure. This would involve a focus on ensuring that a system fails "gracefully" in ways that do not cause harm or jeopardize the larger operation in which it is deployed (Pinelis, 2021).

Another challenge is that, while conventional weapons may feature components from several sources, this is especially true of AI-enabled weapons.[4] This is because much cutting-edge AI development is occurring in the private sector and is being incorporated as components into military systems, and because AI is often utilized to serve specific functions within a larger weapon system. This can make it difficult to assemble large data sets that enable robust tests of all AI-enabled components in a system.

Adapting TEVV to AI-enabled weapons

Given these ways in which AI-enabled weapons are different from conventional ones, the TEVV process needs to be adapted to address the challenges they present. We focus here on key changes that would serve the requirements of *jus ante bellum* to assure safety, precision and accuracy; to avoid unjust resort to war; and to ensure that AI-enabled weapons do not cause unjustifiable harm.

TEVV throughout the weapon lifecycle

The TEVV process should be ongoing. TEVV should track the lifecycle of the system, and some aspects of TEVV need to be repeated when the system gets deployed in a new operational environment (Flournoy et al., 2020). As the parameters of this weapon change in response to different features in its environment, it will be necessary to determine when these changes effectively produce a new weapon that requires a new TEVV, or when a new TEVV is necessary for one or more of its components. Furthermore, it will be necessary for a robust TEVV process not to only assess performance in appropriate operational environments, but to define those environments, often in collaboration with those who are developing or integrating AI into weapon systems.

Training data

Many algorithms relevant to weapon systems, such as object recognition or decision augmentation warfighting algorithms, are trained in simulated environments built on machine-learning algorithms. Simulation-based testing data, however, will be problematic when the risks of deploying a weapon are especially high. In these cases, data sets based on actual conditions are preferable because they can increase commanders' and operators' ability to trust a system in high-risk operational environments.

Gradual deployment

It will also be crucial in many cases that an AI-enabled weapon be deployed only gradually. "[A] strategy of graded autonomy (slowly stepping up the permitted risks of unsupervised tasks, as with medical residents) and limited capability fielding (only initially certifying and enabling a subset of existing capabilities for fielding)

4 Interview with Joe Chapa, Chief Responsible AI Ethics Officer for the Air Force.

could allow the services to get at least some useful functionality into warfighters' hands while continuing the T&E process for features with a higher evidentiary burden" (Wojton et al., 2021, p. 20).

TEVV should consider alternatives to use of AI

While TEVV should be adapted to meet the challenges of AI-enabled weapon systems, it also should be used to help identify when using a human, or some other alternative to AI, should be used for one or more components of a system. This would rest upon assessment of how well different systems would achieve the goals of a weapon, taking into account its performance and risks. In other words, TEVV should not simply assess the safety and precision of a weapon in isolation, but should do so in comparison with available alternatives for similar functions in different operational environments.

TEVV should drive certification schemes

The iterative process used in TEVV can help guide appropriate training, skills and certifications of operators. For example, the US Joint AI Center proposal included four types of testing: algorithmic testing, human-machine testing, systems integration testing, and operational testing with real users in real scenarios (Pinelis, 2021). The human-machine testing and the operational testing provide evidence not just for the evaluation of the weapon, but for how a weapon should incorporate and present machine outputs in order to augment human judgment in the decision-making process in the best possible way. While TEVV has always played a role in US certification schemes for operators, the training content involved in conducting TEVV in certification schemes for AI-enabled weapons may well be significantly greater.

In the ways we have described above, a TEVV process that is sensitive to the challenges of AI-enabled weapons can meet the requirements of the *jus ante bellum*. As the next section discusses, however, this alone will be insufficient to meet these requirements if a state develops these weapons in ways that trigger the security dilemma.

The security dilemma

The security dilemma exists when one state's investment in military capabilities prompts other states to increase their own investments because they perceive that the first state's actions make them less secure. Two factors may be especially important in triggering this dilemma. The first is when states perceive that the offense has the advantage over the defense, and that they may need to act first to preempt a threat. The second is when it is difficult to distinguish whether a state is developing offensive or defensive weapons, which prevents states from signaling their intentions. Together, these can generate a sense of insecurity that creates incentives for states to develop and deploy weapons as soon as possible. As the discussion below describes, various features of AI-enabled weapons may make these conditions especially likely to occur. The result the risk that states could rush to deploy such weapons without conducting rigorous TEVV.

AI-fueled security dilemma

First, AI-enabled weapon systems will not be directly observable in the way that conventional weapons are. Whether a system is enabled by AI depends not upon its visible physical characteristics but the software that guides its operation. This means that it is likely to be extremely difficult for one state to determine the AI-enabled weapon capabilities of another.

Second, the dynamic rate of AI innovation means that even if it were possible to make an assessment of a state's AI-enabled capabilities at one point, this assessment may soon be outdated. Third, AI is not itself a weapon but a technology that can be put to a variety of uses. A state therefore faces a considerable challenge in attempting fully to comprehend all the ways in which other states may be incorporating AI into their military operations. Fourth, unlike during the Cold War, states have little experience with the use of AI-enabled weapons that could provide a shared understanding of their capabilities and risks, and thus a basis for negotiating limitations.

Finally, the nature of AI-enabled weapons may intensify a security dilemma because of the perceived decisive advantage of operating at machine speed compared to a "remotely controlled, 'slower' adversarial system" (Altmann and Sauer, 2017, p. 119). A state may feel especially vulnerable because it fears that another state's use of such weapons against it would inflict grave damage that would prevent it from defending itself or responding. Under these circumstances, states are likely to believe that the balance of military capabilities favors the offense, which can make a preemptive strike seem advantageous.

Avoiding the security dilemma

What might states do to minimize the risk that development of AI-enabled systems will generate a security dilemma that could risk their harmful deployment? One important measure is to avoid using language likely to trigger a sense of insecurity on the part of other states. A state should avoid characterizing its systems as providing it with an unprecedented decisive military advantage over other states. Language that can create the same risk is the public declaration that states are engaged in an "AI arms race." Unfortunately, there is no shortage of such language.[5] Framing the situation in this way suggests that states need to invest in developing and deploying AI-enabled weapons as soon as possible if they want to be secure. A state therefore will need to find a balance between signaling that it has capabilities that should discourage other states from attacking it, while not representing these capabilities as providing it with an overwhelming advantage.

States also can seek to engage in confidence-building measures (CBM) that are designed to reduce states' suspicion of one another through the exchange of information about capabilities and intentions, which may enable some agreement on how operations

5 See, e.g., Geist (2016) and Rickli (2017).

will be conducted.[6] Such measures gained particular prominence during the Cold War as a way of reducing the likelihood that misinterpretation of capabilities and intentions could lead to nuclear war.

One measure is for a state to announce publicly that it is committed to ensuring that deployment of these systems is consistent with ethical principles and legal requirements, and that there is assurance of their reliability and safety.[7] The US Defense Innovation Board, for instance, has released *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, which have been adopted by the Department (US Department of Defense, 2020). These signal to other states that the US military will develop and deploy AI systems only after careful review to ensure that they can be used ethically. In addition, DoD is adapting both its TEVV and weapons review process to conduct assessments of AI-enabled systems. Publicly committing to these measures can serve as a "costly signal" to other states that they will not be disadvantaged by likewise committing to use AI-enabled weapons only after such review.

A second step could be to work to develop common definitions and shared understanding among states of core concepts that are relevant to the safety, reliability, impact, performance, and risks of AI-enabled weapon systems. A third measure would be to encourage information-sharing and communication channels among states. Some degree of transparency about TEVV, for instance, could involve public release of general information about the process for assessment of military AI-enabled systems without disclosing their specific technical features. This would be similar to the US approach to weapons review, which involves disclosing the process but not the review of particular weapons, in an effort to encourage other states to conduct reviews.

States might also share information on how to establish parameters that limit the domain in which a system can operate without human supervision, and how safely to shut it down if it begins to pose risks by operating beyond that domain. There could be some risk to a state from sharing such information, since it could enhance the ability of adversaries to deploy effective and reliable systems that they could use to threaten the sharing state's security. A state therefore would need to decide how to weigh the security risk of an adversary's improved AI capabilities compared to the risk of an adversary and other states deploying unsafe and unreliable AI systems in ethically problematic ways.

The measures described above could also help build confidence by serving as the impetus for a fourth step, which is establishing common norms and codes of conduct about the deployment and use of AI-enabled systems. Over time, states might bolster these measures by taking a fifth step, which is providing for some degree of inspection and verification. One measure could be for states to share the general characteristics of an AI-enabled weapon without revealing all its training data or other components that they may fear would compromise security. Another might be to permit outside parties to observe the operation of the system without disclosing its algorithms.

Finally, states might work to develop "rules of the road" for the conduct of AI-enabled military operations and perhaps "red lines" that establish limits on their use. States also could agree to declare some geographic areas off limits to autonomous systems because of their risk of unanticipated interactions, as well as pledge not to incorporate AI into their nuclear weapon systems.

Conclusion

The concept of *jus ante bellum* expands the just war tradition by suggesting that the way in which states prepare for war can be subject to ethical assessment. The distinctive risks of AI-enabled weapon systems make such an assessment especially important. We argue that ethical development of AI-enabled weapon systems requires that a state engage in rigorous testing of a system before its deployment, and that it develop its systems in ways that do not create a security dilemma that would prompt other states to deploy its own systems without such testing. Both steps can be challenging, but they are essential to ensure that weapons are used in ways that are consistent with human values.

Author contributions

MR took the lead on the section on the security dilemma, while JD did so for the section on testing and evaluation, but each reviewed and helped edit the other's sections. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

6 Among the sources on this subject are Desjardin (2014), Bode and Huelss (2018), Horowitz (2018), Imbrie and Kania (2019), Horowitz et al. (2020), Horowitz and Scharre (2021) and Scharre (2021).

7 It is important to acknowledge that in considering the security dilemma we acknowledge that peer and competitor states might find both generic AI and weapons-specific AI as threatening, but that weapons-specific AI might raise further worries than broad-use AI. The governance of both AI-enabled weapon systems and broad-use AI thus matters for the security dilemma. For governance of broad-use AI, see, for example, White House (2022).

References

Altmann, J., and Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival* 59, 117–142. doi: 10.1080/00396338.2017.1375263

Bode, I., and Huelss, H. (2018). Autonomous weapons systems and changing norms in international relations. *Rev. Int. Stud.* 44, 393–413. doi: 10.1017/S0260210517000614

Boulanin, V., and Verbruggen, M. (2017). *Mapping the Development of Autonomy in Weapon Systems*. Stockholm: Stockholm International Peace Research Institute.

Desjardin, M.-F. (2014). *Rethinking Confidence Building Measures*. New York, NY: Routledge.

Dobos, N. (2020). *Ethics, Security, and the War Machine*. Oxford: Oxford University Press.

Ekelhof, M. (2018). Lifting the fog of war: Autonomous weapons and human control through the lens of targeting. *Naval War Coll. Rev.* 71, 6. Available online at: https://digital-commons.usnwc.edu/nwc-review/vol71/iss3/6/ (accessed May 2, 2023).

Fabre, C. (2021). War, duties to protect, and military abolitionism. *Ethics Int. Affairs* 35, 395–406. doi: 10.1017/S089267942100037X

Flournoy, M., Haines, A., and Chefitz, G. (2020). *Building Trust through Testing*. Washington, DC: WestExec Advisors. Available online at: https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf (accessed May 2, 2023).

Geist, E. M. (2016). It's already too late to stop the AI arms race—We must manage it instead. *Bullet. Atomic Scient.* 72, 318–321. doi: 10.1080/00963402.2016.1216672

Hand, D. J., and Khan, S. (2020). Validating and verifying AI systems. *Patterns* 1, 37. doi: 10.1016/j.patter.2020.100037

Horowitz, M. (2018). Artificial intelligence, international competition, and the balance of power. *Texas Natl. Secur. Rev.* 1, 36–57. doi: 10.15781/T2639KP49

Horowitz, M., Kahn, L., and Mahoney, C. (2020). The future of military applications of artificial intelligence: A role for confidence-building measures? *Orbis* 64, 528–543. doi: 10.1016/j.orbis.2020.08.003

Horowitz, M., and Scharre, P. (2021). *AI and International Stability: Risks and Confidence Building Measures*. Washington, DC: Center for a New American Security. Available online at: https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures (accessed May 2, 2023).

Imbrie, A., and Kania, E. (2019). *AI Safety, Security, and Stability Among Great Powers*. Washington, DC: Center for Security and Emerging Technology. Available online at: https://cset.georgetown.edu/publication/ai-safety-security-and-stability-among-great-powers-options-challenges-and-lessons-learned-for-pragmatic-engagement/ (accessed May 2, 2023).

NSCAI (2021). *Final Report of the National Security Commission on Artificial Intelligence*. Washington, DC: NSCAI (National Security Commission on Artificial Intelligence). Available online at: https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf (accessed May 2, 2023).

Pinelis, Y. (2021). *Presentation on Progress in Testing and Evaluation of AI-enabled Weapons Systems. Video presentation. Hosted by SERC (Systems Engineering Research Center)*. Available online at: https://www.youtube.com/watch?v=1eSKngsJvvo (accessed May 2, 2023).

Rickli, J.-M. (2017). "Artificial intelligence and the future of warfare," in *World Economic Forum, The Global Risks Report, 12th Edn*, 49. Available online at: https://www.weforum.org/reports/the-global-risks-report-2017 (accessed May 2, 2023).

Roff, H. M., and Danks, D. (2018). Trust but verify: The difficulty of trusting autonomous weapons systems. *J. Milit. Ethics* 17, 2–20. doi: 10.1080/15027570.2018.1481907

Scharre, P. (2021). Debunking the AI arms race theory. *Texas Natl. Secur. Rev.* 4, 121–132. doi: 10.26153/tsw/13985

UD Department of Defense (2021). *DoD Instruction 5000.89: Test and Evaluation*. Washington, DC: US Department of Defense. Available online at: https://www.dau.edu/Lists/Events/Attachments/409/DoDI%205000.89%20Test%20and%20Evaluation%208.11.21.pdf (accessed May 2, 2023).

US Department of Defense (2020). *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense. Press Release*. Washington, DC: US Department of Defense. Available online at: https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/ (accessed May 2, 2023).

van der Linden, H. (2010). "Just military preparedness: A new category of just war theory," in *Paper presented at the Department of Philosophy at Michigan State University*. Available online at: https://digitalcommons.butler.edu/facsch_papers/1073 (accessed May 2, 2023).

White House (2022). *Blueprint for an AI Bill of Rights*. Available online at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ (accessed May 2, 2023).

Wojton, H., Porter, D., and Dennis, J. (2021). *Test and Evaluation of AI-Enabled and Autonomous Systems: A Literature Review*. Alexandria, VA: Institute for Defense Analysis. Available online at: https://testscience.org/wp-content/uploads/formidable/20/Autonomy-Lit-Review.pdf (accessed May 2, 2023).

Frontiers in
Big Data

Explores the potential for big data to address global challenges

This innovative journal focuses on the power of big data - its role in machine learning, AI, and data mining, and its practical application from cybersecurity to climate science and public health.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact