Ethics and Information Technology (2024) 26:29
https://doi.org/10.1007/s10676-024-09762-w

ORIGINAL PAPER

# Explainable AI in the military domain

Nathan Gabriel Wood[1]

Accepted: 15 March 2024

**Abstract**

Artificial intelligence (AI) has become nearly ubiquitous in modern society, from components of mobile applications to medical support systems, and everything in between. In societally impactful systems imbued with AI, there has been increasing concern related to opaque AI, that is, artificial intelligence where it is unclear how or why certain decisions are reached. This has led to a recent boom in research on "explainable AI" (XAI), or approaches to making AI more explainable and understandable to human users. In the military domain, numerous bodies have argued that autonomous and AI-enabled weapon systems ought not incorporate unexplainable AI, with the International Committee of the Red Cross and the United States Department of Defense both explicitly including explainability as a relevant factor in the development and use of such systems. In this article, I present a cautiously critical assessment of this view, arguing that explainability will be irrelevant for many current and near-future autonomous systems in the military (which do not incorporate any AI), that it will be trivially incorporated into most military systems which do possess AI (as these generally possess simpler AI systems), and that for those systems with genuinely opaque AI, explainability will prove to be of more limited value than one might imagine. In particular, I argue that explainability, while indeed a virtue in design, is a virtue aimed primarily at designers and troubleshooters of AI-enabled systems, but is far less relevant for users and handlers actually deploying these systems. I further argue that human–machine teaming is a far more important element of responsibly using AI for military purposes, adding that explainability may undermine efforts to improve human–machine teamings by creating a *prima facie* sense that the AI, due to its explainability, may be utilized with little (or less) potential for mistakes. I conclude by clarifying that the arguments are not against XAI in the military, but are instead intended as a caution against over-inflating the value of XAI in this domain, or ignoring the limitations and potential pitfalls of this approach.

**Keywords** Autonomous weapon systems · Artificial intelligence · AI · Explainability · Human–machine interaction

## Introduction

Artificial intelligence (AI) is revolutionizing society, and entire industries are being reshaped in the wake of increased automation and artificial governance. However, though "AI has many constructive applications... [i]t is also being used as a weapon of repression and to gain military advantage".[1] In fact, most militarized states regard autonomous and AI-enabled systems as pivotal technologies in the fight for supremacy in the global order.[2] The incorporation of AI into military systems naturally leads to a host of worries concerning responsibility, predictability, safety, and basic tenets of humanity in war. This is especially the case when such systems are opaque, that is, "the internal factors that determine their decisions are not fully known to people due to the systems' computational complexity".[3] Opaque AI systems are seen to present unique challenges because they undermine human users' abilities to fully understand a system, to follow the processes which led to the system's outputs, and to reliably predict the behaviors of the system. To address these issues, there is a growing body of research on explainable AI (XAI), on "developing approaches to explain

✉ Nathan Gabriel Wood
wood@flu.cas.cz

[1] Institute of Philosophy of the Czech Academy of Sciences, Jilská 1, Praha 1, Czech Republic

[1] Scharre (2023, p. 4).

[2] Scharre (2023) provides extension discussion of the geopolitical battles surrounding AI development and the importance global powers such as the United States, China, and Russia have placed on this technology. See also Horowitz (2020) and Ding and Dafoe (2023).

[3] Peters (2022, p. 963).

Published online: 16 April 2024

and make artificial systems understandable to human stakeholders".[4] This is no less true for AI in the military, and both the International Committee of the Red Cross (ICRC) and the United States Department of Defense (US DoD) have picked out explainability as a key factor in the responsible development and use of autonomous and AI-enabled technologies in war.[5]

In this article, I develop a cautiously critical view of the importance of XAI in the military domain. In particular, I argue that while the methodologies, approaches, and overall goals of XAI point toward clear virtues of engineering and design, these virtues are ones which are not as relevant within the context of contemporary military deployments, many of which will likely see increasing use of autonomous weapons. I further argue that a host of autonomous and AI-enabled technologies used for military purposes fall outside the scope of XAI, due to these systems either not incorporating AI at all, or to them incorporating AI systems that are simple or rudimentary enough that explainability will be trivially present. However, we can expect at least some AI systems which are truly opaque, either due to in-principle limitations to their explainability or to practical limitations in humans that make them unexplainable to us (even though they may theoretically be explainable). For these, I argue that while explainability is a virtue, it is one aimed more toward engineers designing such systems or troubleshooting systems which have exhibited novel unwanted behaviors. However, for the military personnel who must deploy and rely on AI-enabled systems, the ways AI systems are teamed with human combatants will far outweigh any value to be had by explainability. Thus, I argue that for AI in the military domain, the key component for responsibly and safely deploying such systems is that these are integrated into well-established and tightly knit human–machine teams, where the human can reliably predict the AI's behavior and respond accordingly, even when that human does not have a full explanation of the AI's behavior. In developing this point, I draw analogy between sophisticated AI systems and animals fulfilling combat roles, and likewise, explore human–machine teamings through analogy to human-animal teamings. I conclude that XAI does have a role in military affairs, but maintain that this role is related primarily to the development and troubleshooting of AI systems, and has less role in actual deployments of AI in military contexts.

The arguments are structured as follows. First, I begin (Sect. 2) by clarifying a number of key definitional points. With these in place, I examine simple autonomous and AI-enabled systems which are currently in use in the military or will be in the near future (Sect. 3). In canvassing such existing and near-future systems, I highlight that XAI plays little role in the military systems of today, due to the relative transparency of AI processes in these systems. Yet though current military systems have simpler or more transparent AI systems, this will not always be the case, and in Sect. 4 I continue by examining the role XAI may play for more distant AI systems in the military. In exploring this, I examine how opaque AI can contribute to unpredictability in systems (Sect. 4.1) and I compare the values offered by XAI against those to be gained through a richer implementation of human–machine teaming in the military (Sect. 4.2). In these discussions, I emphasize that there are limitations to the practical value of explanations, and highlight that the value will vary depending on where XAI is implemented and for whom. Finally, I conclude (Sect. 5) by reiterating that XAI does have value in the military domain, but that this value is not one primarily related to responsible *deployments* of AI, but rather to responsible innovation and design of these systems and effective troubleshooting of systems which exhibit novel unwanted behaviors.

## Autonomous weapons, AI in the military, and explainability

Before beginning any discussion of AI, autonomous weapons, or opaque systems in the military, it is crucial that the exact understanding of these terms be made explicit at the outset, as "underdeveloped or underclarified view[s] can, and most likely will, lead to confusion, error, and much time and effort squandered".[6] This is especially the case for emerging technologies, where there are likely to be many competing definitions, each of which holds some merit. This section will thus be devoted to providing brief explications of what I mean in this article by "autonomous weapon system", "human–machine teaming", "artificial intelligence", "opacity", and "explainable AI". However, it is worth stressing that I am not arguing for the definitions or understandings provided (as there is reasonable room for disagreement), and instead am merely clarifying the meaning of the terms as they will be used throughout what follows.

Now, as many debates surrounding AI in the military focus on autonomous weapon systems (AWS), we will begin with these. In the past decade and a half, there have been many definitions of AWS provided by scholars, states, and non-governmental organizations.[7] However, there

[4] Langer et al. (2021, p. 1). See also Miller (2019) and Mittelstadt et al. (2019).

[5] See, respectively, International Committee of the Red Cross (2021a, p. 7), US Department of Defense (2023, pp. 4, 6).

[6] Wood (2023a, p. 10).

[7] See Williams (2015), Boothby (2016), Altmann and Sauer (2017), Caron (2020), Taddeo and Blanchard (2022) for overviews and taxonomical work. See also Pacholska (2024) for discussion of subtle dif-

is increasing acceptance of the definition put forward by both the ICRC and the US DoD, namely that AWS are to be understood as weapon systems that have autonomy in the "critical functions" required for selecting and engaging targets,[8] and that they can select and engage targets without human intervention.[9] This definition captures the essential features of autonomous weapon systems, namely that they are autonomous in their core tasks, but it does not imply that such systems possess any sophisticated internal AI programming, nor that they are opaque, unpredictable, or even necessarily lethal. In fact, under this definition, there are many AWS which have been in use around the world for decades, from anti-radiation missiles to close-in weapon systems, as well as many others.[10]

In evaluating the impact of any of these systems, it is also critical to look not just to the capabilities and limitations of the systems themselves, but to also pay heed to how these systems are integrated with humans into cohesive units. This is what is known as human–machine teaming, and pertains to every technology in war. At the upper end, we might think of systems like unmanned aircraft which can carry out complex tasks autonomously, even selecting and engaging targets, but which have humans overseeing them and giving the green light on distinct engagement decisions. In this type of teaming, the human must understand the system, its capabilities and limitations, and the engagement context well enough to competently gauge the reliability of the system and halt it if necessary. But human–machine teamings go all the way down to the lowest tech items in war as well. Recall the words of the Rifleman's Creed of the United States Marine Corps:

> This is my rifle. There are many like it, but this one is mine.
> My rifle is my best friend. It is my life. I must master it as I must master my life.
> Without me, my rifle is useless. Without my rifle, I am useless.

For any technological system in war, even a rifle, its capacity to provide advantage is deeply entwined with its integration into capable and reliable human–machine systems (or perhaps human-artifact systems, for simpler things like firearms). More than this, responsible use of any technological system demands that the humans making use of these have a sufficient understanding of the system itself. This is central to human–machine teaming.[11]

Returning to autonomous weapons and artificial intelligence, while it is true that many of the AWS currently in use utilize little to no AI, or have only rudimentary AI systems enabled, this is already and rapidly changing. As such, it is also critical that we are clear about precisely what we mean by "artificial intelligence". Following some of the pioneers of AI research, we may with our definition "wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity",[12] or that we are "concerned with methods of achieving goals in situations in which the information available has a certain complex character".[13] These notions are somewhat vague though, and for the sake of precision I follow (Wang, 2019), taking for granted that

> [i]ntelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources.[14]

The degree to which a military system possesses "AI" will thus be determined by that system's capacity for adapting to its environment given insufficient data. The more a system is able to accomplish goals and secure gains while operating under such limited conditions, the more strongly we may maintain that it is AI-enabled. And since military systems will, as a rule, usually be operating with limited information and resources, there will be pressure to develop more and more sophisticated AI systems, even when this entails that such systems may by necessity be less transparent or understandable. Which brings us to opacity and the push for explainable AI.

Footnote 7 (continued)

ferences between certain core definitions from states and non-state actors.

[8] International Committee of the Red Cross (2014, p. 5).

[9] International Committee of the Red Cross (2021b, p. 1), US Department of Defense (2023, p. 21).

[10] Boulanin et al. (2020), International Committee of the Red Cross (2021a), Heller (2023), Wood (2023a), and Wood (2023b).

[11] Human–machine teaming, the ways it may be pursued, and XAI generally have direct and important implications for the idea of "meaningful human control" (MHC) of AWS, a guiding principle which has become central in many debates on autonomy in military systems. However, though there are clear touch-points between these, the depth and breadth of the discussions of MHC makes it impracticable to explore these within the context of this work. For discussion of MHC at a general level and specifically with regards to AWS, see respectively, e.g., Santoni de Sio and van den Hoven (2018), Mecacci and Santoni de Sio (2019), Ekelhof (2019), Human Rights Watch (2016), and Bode and Watts (2021).

[12] Newell and Simon (1976, p. 116).

[13] McCarthy (1988, p. 308). See also Minsky (1985).

[14] Wang (2019, p. 19). See also Wang (1995) and the 2020 special issue of the *Journal of Artificial General Intelligence* dedicated to discussing Wang's view (Volume 11, Issue 2). For a slightly more technical definition from the law, see the EU AI Act, esp. p. 39.

As AI systems become more complex, it becomes increasingly difficult for humans to be able to fully comprehend, understand, or explain how they function. This may be due to simple practical limitations (e.g., the AI makes use of too many interconnected functions and algorithms for a human to feasibly be able to parse the code, even if it is in principle possible) or be the result of genuine barriers to understanding (e.g., the AI makes use of machine learning approaches or deep neural networks which prevent a human from being able to understand the underlying reasoning processes). In such cases, we may consider these systems to be *opaque*, or to use alternative terminology, we may call such a system "a 'black box'... a system for which we know the inputs and outputs but can't see the process by which it turns the former into the latter".[15] In the military domain, such "black boxes" would appear to present a uniquely thorny problem, and it is unsurprising that XAI efforts were spearheaded by military researchers, with the growing visibility of this research owing much to projects run by the United States Defense Advanced Research Projects Agency (DARPA).[16]

In order to remedy these difficulties, XAI seeks to "(1) produce more explainable models while maintaining a high level of learning performance (e.g., prediction accuracy), and (2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners".[17] More simply, "[t]he purpose of an explainable AI (XAI) system is to make its behavior more intelligible to humans by providing explanations".[18] Given the recent boom in research on XAI, a number of approaches and methods have been proposed,[19] but in general all methodologies will be aiming toward some version of goals 1) and 2) above. In the military domain, this is no different, as those designing and deploying potentially opaque AI systems will always be balancing the military advantages of speed and precision against the moral and legal need to have systems which are both predictable and sufficiently understandable to the combatants making use of these technologies.

As a final point, it is worth making clear that throughout the arguments to come, I am assuming that the actors involved in the development and deployment of AI in the military are (at least) trying to act in good faith and in the spirit of the ethics and laws of war. At a minimum, I assume that such good faith requires efforts to adhere to Article 36 of Geneva Protocol I Additional to the Geneva Conventions, namely that in the development or adoption of a new weapon parties try to "determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law". And following on this, given the assumption of actors acting in good faith, we can further assume that the programming of AI systems in the military domain will in general and by default be set to conservative targeting parameters; i.e., autonomous and AI-enabled systems will be designed so as to aim to minimize false positives in targeting, taking as a cost an expected increase in false negatives.[20]

With this rough definitional groundwork laid, we can now move onto the arguments. However, before doing so, it should again be noted that I am not arguing that any of the above definitions or understandings ought to be considered *the* definition or in some sense "better than" alternative views. I have opted for definitions in keeping with either a broad selection of scholarship or reflecting the views of central state and non-governmental organizations, but there is merit in probing alternative definitions and their implications. For the purposes of this article, we will move forward with the understandings just sketched, but one may reasonably examine these topics through other lenses as well.

## Rudimentary AWS and AI

Advanced militaries have long had access to autonomous weapons and systems enabled with at least rudimentary forms of AI. Importantly, the vast majority of autonomous weapons currently in use are either advanced autonomous munitions or anti-materiel platforms which operate based on rather clear targeting parameters.[21] For these, the question of explainability is moot, as such systems are in most cases not utilizing AI of any sort. Rather, anti-radiation missiles locate and engage objects emitting radio signatures associated with radar stations and jammers, anti-tank munitions utilize seismic, acoustic, or high-frequency radar to track

[15] Michel (2020, p. iii).

[16] Adadi and Berrada (2018, p. 52144). See also Gunning et al. (2019), Gunning and Aha (2019), and Gunning et al. (2021) for discussion specific to the DARPA project on XAI and Michel (2020) for overarching assessments of explainability and predictability in military systems.

[17] Arrieta et al. (2020, p. 83).

[18] Gunning et al. (2019, p. 1). For the role of social and ethical considerations in XAI methodologies, see, e.g., Miller (2019), Langer et al. (2021), and Peters (2022).

[19] For surveys and taxonomical discussions on the state of the art, see Adadi and Berrada (2018), Das and Rad (2020), Arrieta et al. (2020), Fiok et al. (2021), Speith (2022), and Cambria et al. (2023). See Ross (2022) for critical remarks on the push for transparency.

[20] Article 50.1 of Additional Protocol I stipulates that "in case of doubt whether a person is a civilian, that person shall be considered to be a civilian", setting the basic justification for such conservative targeting parameters. Many thanks to Maciej Zając for suggesting this clarification.

[21] Wood (2023b, pp. 4–10).

heavy vehicles and armor, and close-in weapon systems used for missile defense take primarily speed and heading of aircraft as parameters to determine whether or not something is a threat. In these and other similar systems, many of which have been in use for decades, AI is not necessary, and is usually not present (except perhaps for limited purposes). As such, explainability holds no particular relevance for AWS *per se*. Rather, the critical value is *predictability*; if a combatant can reliably predict how an AWS will function in the contexts where they plan to deploy it, then that may suffice for responsibly and safely utilizing such systems. Moreover, if a simpler AWS is predictable, there is no clear reason why the combatant deploying it would need to be able to explain its "actions". Knowing when it will function correctly and when it won't, and responding accordingly to that knowledge, will suffice for the ethical and legal use of these systems. Knowing *why* the system makes certain targeting decisions in certain contexts might help combatants to more quickly grasp the "dos and don'ts" of deploying AWS, but one need not be an engineer or programmer in order to recognize that some battlefield situation is one which is likely to cause an accident. After all, training in the use of weapon systems is meant to teach combatants when such systems will and won't work, but knowing this does not require the users of those systems to be troubleshooters, repairmen, and designers of those systems as well. At any rate, there are a host of autonomous weapons with no AI, and for these, XAI is not needed.

However, in addition to simpler AWS, we are seeing increasing development toward advanced AI-enabled systems which can operate with far less human oversight, and which can accomplish far more complex tasks. In fact, the major global powers are increasingly locked in what might be seen as an arms race for AI, with Vladimir Putin claiming that "[a]rtificial intelligence is the future... [and] whoever becomes the leader in this sphere will become the ruler of the world", Xi Xinping adding that "[s]cience and technology has become the main battleground of global power rivalry".[22] In response to such positions, the United States has recently announced plans to begin using thousands of new autonomous systems over the next 2 years.[23]

Yet statements on the importance of AI, and even plans to utilize a greater number and variety of AWS do not imply that AI in the military will rapidly be dominated by opaque "black box" systems. Rather, for those current and near-future systems which do have some form of AI (which is only a portion of all autonomous systems in the military), many of these utilize more rudimentary programs which are (likely to be) transparent. And for those with aspects which may be opaque, these often relate to unobjectionable applications of AI. For example, DARPA's Air Combat Evolution program has made use of many recent breakthroughs in AI research to develop autonomous AI systems capable of effectively engaging in dogfights (close-range air-to-air engagements). In a recent competition, the AlphaDogfight Trials, AI pilots were even able to reliably outperform humans in a number of areas.[24] These AI systems are highly complex, and due to how they were designed and trained, are almost certain to be opaque with regards to a number of decisions they may make. However, in the arena of air-to-air combat between jet fighters, there is far less likelihood of targeting mistakes or novel unwanted or ethically suspect behaviors developing. Thus, even when opacity comes into the equation, this will not automatically imply that there is a problem, nor that XAI is needed. Explanations may prove *useful* for a variety of reasons, but their having value does not indicate that their lack speaks against a certain AI system or its deployment to theaters of combat.

## Advanced AWS and AI in the military

Rudimentary AWS often possess no AI, and for many of those that do, the AI is simple or straightforward enough to be transparent by default. And for many current AI systems in the military which are opaque, their opacity does not necessarily undermine their ethical or legal permissibility (as the opacity may only impinge on ethically neutral decisions or decisions where mistakes are extremely unlikely by default). However, as AI continues to improve, continues to be applied to a greater array of tasks, and continues to become increasingly complex (and likewise, opaque), it may begin to appear necessary that XAI be treated as a basic requirement for responsibly utilizing AI systems. In this section, I resist this broad conclusion. In particular, I argue that XAI will often be irrelevant to responsible *deployments* of AI (though it will likely have value at other stages of an AI's design- and life-cycle), that rich and deeply integrated human–machine teamings present a much stronger method for mitigating the possible negative consequences of opacity, and that XAI may even undermine responsible deployments by serving as a form of "check box" for permissibility and thus reducing the impetus for strong human–machine teams.

[22] Quotations found in Scharre (2023, p. 9). See Hunter and Bowen (2023) for critique of the AI hype in the defense domain.

[23] Layton (2023).

[24] DeMay et al. (2022), and Scharre (2023, pp. 1–3).

## Unpredictable AWS and opaque AI

AI systems may be practically opaque in virtue of the sheer number and complexity of (interrelated) functions and algorithms operating in their background. Additionally, autonomous weapons or AI systems which are designed around deep neural networks (or which, more broadly, make use of machine learning for their training) are apt to be in principle opaque due to the fact that a designer or engineer cannot fully track what the system has learned and how it has gone from training inputs and operational data to discrete outputs. Some authors further argue that machine learning not only impacts on the opacity of a system, but will in fact make AI-enabled systems inherently unpredictable as well.[25] If we return to the definition of intelligence presented in Sect. 2 above, we can see why this may indeed be the case.

> Intelligence is the capacity of an information-processing system to *adapt* to its environment while operating with *insufficient knowledge and resources*.[26]

If we understand "artificially intelligent" systems in the above manner, it is to be expected that these will have some capacity for acting in ways which we would deem unpredictable. This is because such systems will need to be trained on massive data samplings in order to be at all effective or to be responsibly deployed. However, that training will inevitably not include every possible scenario they may encounter, or at least not include every scenario from every angle, in every environment, in every type of weather, etc. Quite simply, the system will need to be trained to a sufficient degree of robustness, but it will still have to make calls during actual deployments which are made against a backdrop of incomplete information or information which it has not directly encountered during training. In this way, such systems will almost always have some inherent capacity to surprise us, simply because we cannot have trained them for everything, and when they come across some novel scenario (or a previously encountered scenario, but from a new angle), they may act in novel ways. Importantly, this is not to say they must have *in situ*, or real-time machine learning capabilities, as this can lead to much deeper types of unpredictability and significant challenges for responsibly deploying such systems.[27] However, systems must be able to, in keeping with the training they have received, act in partially novel ways to achieve goals in not only environments their trainers have foreseen, but also environments and contexts that may involve unanticipated variables. Such adaptive problem solving may moreover sometimes lead to behaviors which we cannot fully predict. At least, this much seems plausible. However, the fact that one cannot fully predict certain behavior does not imply that this behavior is unpredictable (in some troubling sense). To see this, let us consider Holland Michel's words on predictability presented in a recent report of the United Nations Institute for Disarmament Research (UNIDIR).

> All autonomous systems exhibit a degree of inherent operational unpredictability, even if they do not fail or the outcomes of their individual action can be reasonably anticipated. This is because, by design, such systems will navigate situations that the operators cannot anticipate. Consider a fully autonomous drone that maps the interior of a network of tunnels. Even if the drone exhibits a high degree of technical predictability and exceptional reliability, those deploying the drone cannot possibly anticipate exactly what it will encounter inside the tunnels, and therefore they will not know in advance what exact actions the drone will take.[28]

Michel is correct in pointing out that one cannot "know in advance what exact actions the drone will take", especially when one is considering systems with opaque architectures. However, the same is true of human combatants sent to carry out similar missions. In fact, if we consider fully determinate computer systems, where each input has a clear unique output, it is also the case that for these we cannot know in advance exactly what they will do. This is because we cannot know in advance what they will encounter. But even though we do not know *exactly what they will do*, we do know what they will do *given certain situations*. The same is true, though to a lesser degree, for human combatants sent on missions like the one Michel imagines. The question thus should not be whether we can predict what will happen, but rather whether we can predict what will happen *given various inputs*. For the sake of argument, let us assume that opacity alone undermines our ability to do this, to reliably predict what will happen given particular inputs.[29] Would

[25] Blanchard and Taddeo (2022). Note that Blanchard and Taddeo are utilizing a rather stringent definition of AWS, which greatly impacts on their arguments. For fuller discussion of their definition, see Taddeo and Blanchard (2022).

[26] Wang (2019, p. 19), emphasis added.

[27] This is precisely the objection laid out in Blanchard and Taddeo (2022). Haugh et al. (2018) and Verbruggen (2022) present additional concerns relating to the testing and evaluation of autonomous and AI-enabled systems, and McFarland and Assaad (2023) discusses the legal challenges in weapons review raised by *in situ* learning. However, McFarland and Assaad, while indicating a number of complications raised by such online learning, do not argue that this by default or necessity renders such weapons inherently unpredictable or illegal to use. Rather, real-time learning alters the necessary review process for weapons with this capability, making it far more stringent. At any rate, to simplify the arguments developed here, I am assuming AI systems which do not make use of *in situ* learning.

[28] Michel (2020, p. 5).

[29] Note that this need not necessarily be the case, or may potentially be mitigated through extensive training, testing, and evaluation. For discussion of ways in which training, testing, and evaluation may raise our confidence in opaque systems, see Zając (*unpublished manuscript*).

XAI greatly improve the situation or remove this element of unpredictability?

In order to answer this, we must first differentiate between systems which are truly autonomous and will be deployed without contemporaneous human oversight of any kind (human off-the-loop), those where the system functions autonomously but can have its decisions overridden by a human (human on-the-loop), and those where a human at least partially controls (some of) the system's functions and targeting decisions (human in-the-loop). Looking first to off-the-loop AWS and AI-enabled systems, we will see that XAI can have no real role during deployments of these.

If we are envisioning truly *autonomous* weapon systems imbued with opaque AI, these will be carrying out missions without any contemporaneous human oversight.[30] Designing these systems to provide intelligible and helpful explanations for every decision taken can greatly facilitate the speedy and effective training of such systems, and in the event that a system makes a mistake or does some novel and unwanted thing, provisioning of its "reasoning" will likewise streamline the troubleshooting process. However, for AI systems operating without human oversight, explanations hold zero value during deployments. More than this, it is not possible to have a useful review of explanations pre-deployment as a sort of "check" on the system's expected reliability. This is due, first and foremost, to Michel's concerns about predictability just discussed; an AI system may possess the capacity to provide explanations for its actions, even *ex ante*, but one cannot know in advance exactly what the system will encounter during deployment, or even if one can know this, one cannot know the precise details of how particular objects or targets will be encountered (the angles of approach, ambient temperatures, visual and other lighting of the objects, etc.). These factors are all apt to be highly relevant for the machine's decision-making processes, and the only possible sort of explanation that could be given *ex ante* thus would be an unwieldy listing of factors which may be relevant and may be encountered. Such a list will invariably include too many items to present a useful aid to humans pre-deployment, or it will need to be trimmed and curated, leaving off potential constellations of input data which might impact on the decisions reached. In short, for systems acting without contemporaneous human oversight, explanations before the fact will almost certainly be either too numerous to prove useful or be limited but not fully representative of what the machine may encounter (or some combination of both). And even if these issues can be surmounted, there is the fundamental obstacle that off-the-loop systems have no one to review the decisions while the machine is in operation (though they may before or after deployment). As such, while XAI may improve the pre- or post-deployment development and troubleshooting of such AWS, it will not provide a useful tool for these *during deployment*.

What of systems where humans are on- or in-the-loop? If humans can override the machine's decisions or are part of that decision-making process, it would seem that explanations, especially intelligible ones, could help us to more predictably, reliably, and responsibly use such systems. However, before we become too enamored by this possibility, we have a responsibility to grapple with the challenges associated with XAI and the risks it may bring when deploying AI in military contexts. The remainder of this and the following subsection will be devoted to examining some of these risks and challenges.

The first area of potential worry is the design of XAI systems, and whether the explanations provided are actually doing any good for combatants responsible for overseeing AI systems deployed to combat environments. This is a significant area where care is required, as poor explanations or explanations which do not highlight the right factors underpinning the AI's evaluation are apt to lead to mistakes. For example, Rudin (2019) presents the case of an AI system tasked with identifying images, and shows how faulty "explanations" may lead to confusion and over- or under-confidence in systems. In point of fact, Rudin's example centers around an image of a dog and two accompanying heatmaps showing the points the AI found relevant for two separate identifications of the image. Both heatmaps are remarkably similar, but one is explaining what points the AI system found relevant for its assessment of "Evidence for animal being a Siberian husky", whereas the second shows the points relevant for "Evidence for animal being a transverse flute".[31] The similarity of "explanation" for these wildly divergent assessments indicate just how flawed and misleading explanations can be.

This is especially problematic given the tempo of modern warfare and the need for overseers of AI systems to make

[30] Some argue that such systems present ethical and legal challenges of their own, given that these necessitate that decisions to potentially kill human beings will be delegated to machines. Going into these debates here is beyond the scope of this article, but it is worth mentioning that there already exist many autonomous systems that can carry out lethal engagements without human oversight. For example, many missile defense systems are designed to intercept both incoming missiles and high-speed aircraft, the latter of which engagements may obviously be lethal. These are, however, routinely not subjected to outcry or objections from AWS critics. This is arguably due to the necessarily defensive nature of such systems, but it highlights that neither complete autonomy nor lethality are at the root of objections to AWS. Nor indeed can the delegation of life-and-death decisions to machines be the essential objection, otherwise these systems would also see this form of critique (which they do not).

[31] Rudin (2019, p. 209). See also Ch. 5 of Michel (2020) for further discussion of this case and discussion of problems with XAI.

rapid decisions. If a combatant has seen the AI-enabled system perform well across a variety of contexts, and has always associated the explanations given with something akin to justifications for targeting decisions, then it is entirely possible that flawed explanations may not be easily or reliably noted. More than this, explanations which highlight the wrong elements or do not include the aspects which the AI is apt to misidentify may fail to give combatants any significant opportunity to confidently intervene when necessary. This is not to say that explanations necessarily will be flawed in this way or cannot be done well, but merely to indicate that XAI can create serious risks if executed poorly.

One may hope to mitigate the above worry by including explanations that are richer, or which highlight what factors are included in the explanation, which are excluded, and what weightings are placed on various input data. However, just as explanations which provide too little (or unhelpful) information may cause problems, so too will those which present more than is necessary. First, there is the obvious problem that modern warfare places combatants under increasingly strict time constraints, limiting their ability to engage with lengthy and involved explanations. Moreover, there is the added difficulty that explanations which are rich enough to clarify the underlying problems that may be lurking in the machine's reasoning processes are likely to be complex, delve into aspects of the system's programming and training, or require presentation of large amounts of factors (as many details will likely go into every decision made by the AI system). These may prove to demand more of combatants than is reasonable, requiring that deployers of AI systems be trained as computer scientists and engineers, in addition to their training as warfighters.[32] At any rate, XAI will, by necessity, have to strike a balance between too much and too little in explanations, as either end of the spectrum brings risks of its own.

These are design problems though, and perhaps we can reasonably assume that these will be addressed in time. Even so, the inclusion of XAI during deployments of AI systems is apt to create further obstacles to responsible use of such systems. The primary issue is that the provisioning of rich, intelligible, and informative explanations may give rise to the perception that AI systems may be deployed with more ease or with a possibility of having generally trained users which can reliably and responsibly handle a variety of such systems.

There are two distinct issues at play here. The first is that the presence of XAI may give a perception that humans trained on similar systems (but not the exact system to be deployed) can reliably utilize other systems. The presence of explanations for action, coupled with a humans' training on AI systems generally, may lead to a belief that one can swap between systems with relative ease. However, opaque systems, even ones which give explanations for their actions, are apt to have many subtle factors which go into each decision. These subtle factors may not always be present in explanations, and are in fact likely to not be present if explanations are compact and simple enough to be usable during combat. As such, understanding *these* and responding to them will require that handlers of such systems are deeply familiar *with the particular systems being deployed*. However, XAI may lead to a perception that "one training fits all", undermining the human–machine teamings necessary for responsible deployment.

Second, on a related point, XAI may also lead to a perception that humans may simply "operate" AI-enabled systems without needing to be teamed with them in a rich way at all. This is because the presence of rich and informative explanations may lead to a belief or general sense that anyone can utilize the system so long as they are engaging with the explanations in a critical and thoughtful manner and understand the system and warfighting context well enough to intervene when the system is going to make a mistake. However, as above, the explanations provided are very unlikely to include *all* of the subtle factors and cues which underpin a specific engagement decision. Moreover, the ability to grapple with the subtleties of a particular AI system will likely require that a human have somewhat intimate and firsthand knowledge of that system's functioning. This is likely to only be accessible to humans through their incorporation into rich teamings of humans with machines (ideally, involving cooperative training of both the system and human together). By deploying AI systems which are explainable but are under the purview of those who are uninitiated (or poorly initiated), we would create significant risks for mistake simply in virtue of the fact that "users" of those systems would not possess the relevant knowledge to know which explanations may themselves be suspect, or which might require additional scrutiny.

All of that being said, XAI clearly does have value for military uses of artificial intelligence. However, that value is primarily one related to design and troubleshooting. Knowing the reasons an AI system has for some action can greatly help engineers and programmers in developing systems that are responding correctly to information gathered about their environment, that are giving conservative targeting selections, and that are acting in accordance with the legal and moral requirements of war. In a similar vein, if an AI system makes a mistake during a deployment or begins to display novel and unwanted behaviors, explainability can represent significant value by making the troubleshooting

[32] Additional data coupled with the tempo of warfare is also apt to lead to information overload, nullifying any gain had by the explanations themselves. For general discussion of this issue, see, e.g., Buchanan and Kock (2001) and Phillips-Wren and Adya (2020).

process much quicker, simpler, and more effective; the more clearly an AI system can identify and communicate its reasoning for some action, the better engineers, programmers, and machine trainers can address whatever aspects of its programming or training led it to carry out the unwanted action. These are all ways in which XAI can promote both the development and improvement of AI systems used in the military domain.

However, these are tasks related to the pre- and post-deployment phases, and do not indicate that XAI greatly contributes to the responsible use of AI in discrete military applications. Moreover, the arguments above indicate that XAI will often be irrelevant *during* engagements, and could even be counter-productive. The core problem is that XAI, if successful, will provide more information to combatants, but it will not necessarily imply that said information is well utilized. More importantly, XAI has no innate or necessary connection to human–machine teaming, given that humans may be paired with systems and given adequate training without necessarily having a deep understanding of exactly why a system does what it does. Moreover, that human–machine teaming is a central factor for responsibly using AI in the military domain, and while it is possible that XAI might supplement these teamings and improve how well combatants can deploy advanced artificial intelligence on the battlefield, critically, such success will depend first and foremost on the teamings themselves, and will, at best, be further aided by XAI, at worst, undermined by it. We should therefore be cautious in our optimism about the benefits of explainability for combatants deploying AI for warfighting purposes.

## Human–machine teaming

For autonomous weapons and AI systems which are opaque and potentially unpredictable, explanations may help in designing these systems better or improving those which show faults, but they are unlikely to mitigate the negative effects of opacity and unpredictability during actual military uses of these systems. Moreover, rich and informative explanations may undermine the perceived need for strong human–machine teams, and it is these which are most crucial for reliable, predictable, and responsible uses of AI in the military. In particular, we must ensure that we will have human–machine teams developed from training of AI systems up through their deployments, and with an eye to having dedicated handlers responsible for individual AI-enabled combat systems (or possibly small groups of interlinked systems).

Building on the arguments developed in Wood (2023b),[33] the first point worth stressing is that for opaque AI systems, we ought to recast our thinking about how we engage with these. In particular, we ought to dispense with the language of humans as "users" of these systems, and instead view humans as "deployers", or, better yet, "handlers" of AI-enabled systems. Further still, we should conceptualize an AI system's actions and our impact on them as relevantly analogous not to those of other technical artifacts, but rather to animals' actions.[34] The reasons for this are many, but let us briefly canvas the main points.

If we are assuming that actors are acting in good faith, opaque AI systems used in the military will not simply be built and then deployed. Rather, they will undergo extensive training which familiarizes them with the greatest possible array of situations and complicating factors. They will also be tested against a large variety of combat situations, in contexts where certain variables are apt to lead to errors or mistakes. In point of fact, responsible developers will "look for problems as hard as they can *and then find solutions*".[35] All of this will result in systems which, while still potentially opaque, behave in predictable ways across a large number of contexts. However, despite our ability to generally predict their behavior, that opacity, coupled with the system's own inbuilt capacity for autonomous action, will mean that AI-enabled systems can act in wholly unpredictable ways. In other words, responsibly developed AI systems will be generally predictable, but capable of acting unpredictably.

This is the same situation for animal combatants used in war. Animals have long been a part of mankind's warfare, fulfilling a wide variety of roles,[36] but for the sake of specificity, we may imagine an opaque military AI system as analogous to a combat assault dog.[37] Such dogs are given extensive training, teamed with a human who understands

[33] See also Roff and Danks (2018) and Baker (2022) for similar points and analogies.

[34] Now, there are obviously many *dis*analogies between animals and AI-enabled autonomous systems, but there are deep analogies as well, and ones which are central to the debates here; with sufficient training and testing, both act predictably and reliably; despite that, both *can* act unpredictably in certain contexts; both are opaque to those handling them; both create possible uncertainties about responsibility for the outcomes brought about by their actions (i.e., who is to be held responsible if something goes wrong); and finally, both occupy an uncertain moral and legal space in warfare. For further discussion of these analogous and disanalogous aspects, see Crootof (2018, pp. 76–78), Wood (2023b, pp. 10–12). See also Flemisch et al. (2003) for useful broader discussion of analogies for autonomous systems.

[35] Wood (2023c).

[36] See Nowrot (2015).

[37] Though animals have been increasingly phased out of most functions they previously fulfilled, there are still certain animals, dogs in particular, that continue to work alongside human combatants, sometimes in combat roles. See Baker (2022, pp. 16–19).

them extremely well, and put into combat situations to carry out certain tasks that humans cannot, or that humans cannot do as well as the dog could. Importantly, due to the amount and quality of training they receive, as well as the quality of their teaming with a human, combat assault dogs are generally very predictable. Yet even so, they are still autonomous, and can act in novel and sometimes unwanted ways. It is the responsibility of their human handler to recognize situations where the dog is apt to act unpredictability (for whatever reason), and to respond accordingly. And though there is a gap in the law regarding animal combatants,[38] it is reasonable to hold the handlers responsible in the event that mistakes are made.[39]

Connecting this to the discussion of XAI, human handlers responsible for animal combatants will generally have a strong understanding of when their four-footed friends may be expected to behave normally and when they may be unpredictable. Yet an animal's mind is not something that can be accessed, and it is not possible for handlers to extricate the exact reasons for their charges' actions. Quite simply, animals are opaque. This opacity does not mean that they are wholly unpredictable though, nor even that they are generally unpredictable, or prone to unpredictable action at all. But critically, the predictability of an animal combatant has much to do with *who is doing the predictions*.

Handlers responsible for animals may be extremely reliable predictors of the animals' actions, while other combatants may have no idea at all. Additionally, one's general understanding of the underlying reasons for some animal's actions may also not provide strong predictive reliability. Thus, an animal psychologist may be able to say what drives dogs in general, what reasons they might have for certain actions, and even what may drive particular dogs in combat situations. However, the psychologist looking from the outside is likely to be a far worse predictor of some dog's actions than its handler would be. And this is apt to be the case even if the psychologist has some deeper understanding of the underlying reasons driving the animal; familiarity and mutual trust simply provide far more than mere explanations ever could. And finally, there is the critical point that not only will handlers know when animals may be unpredictable (in potentially unwanted ways), but also when they will be predictably misbehaved. Predictable misbehavior is a key limitation of where and when autonomous agents, organic or otherwise, may be deployed, and knowing when this is likely is best achieved through rich teamings of humans and other agents. Moreover, provisioning of explanations to individuals who are otherwise unfamiliar with an agent, be it a dog or AI, is unlikely to suddenly impart the necessary general understanding required for responsible deployment of such subordinate agents. To see this, consider an example.

> *Buddy*: I have a dog who I take for a walk every day (his name is Buddy, and he is a good boy). As his owner (and handler) I know him very well, to the point that I can reliably recognize (at least) six distinct forms of sniffing he may exhibit: (1) sniffing to just generally engage with the world, (2) sniffing to find a place to go to the bathroom, (3) sniffing because a lady dog came by recently, (4) sniffing because he thinks there might be food, (5) sniffing because he *knows* there is food and he is trying to find it before I stop him, and (6) sniffing because there is something disgusting he would like to roll in.

Each of these forms of sniffing is rather distinct and can be easily distinguished from the other. Moreover, the different types of sniffing result in different actions I might or must take. If he is looking for a place to go to the bathroom, I should bring him to a patch of grass. If he is aimlessly looking for food, it may be prudent to put him on the leash (though that is not necessary). If he clearly knows food is near and is trying to find it before I do, I have a responsibility to put him on the leash immediately (some common food items we eat can be lethally poisonous to dogs). At any rate, it is clear that why he is sniffing impacts on what responsibilities I have. Moreover, these types of sniffing make him predictable. However, and critically, he is predictable *to me* (and my wife). Another individual without deep familiarity with Buddy will simply see a dog sniffing. More than this, I could provide detailed explanations of what each type of sniffing looks like, what they mean, and what responses the human should undertake. However, even these are apt to be unhelpful. After all, his sniffing is a bit faster and more frantic if he's sniffing because he knows there is food. But to the uninitiated, the natural question is "Faster and more frantic *than what*?". Without knowing him already, without having a baseline of understanding concerning his usual behavior, what markers he presents, and what factors are relevant, the explanation provides little. More than this, there are with certainty a number of visual and other cues which I take note of but which I cannot fully explain myself. In point of fact, humans are opaque, and our opacity means that we cannot fully understand exactly why we sometimes know that certain agents will or won't act in certain ways. Quite simply, familiarity breeds a sort of understanding that mere explanations cannot capture, and we ignore that to our peril. And this is true whether the familiarity is with an animal or an artifact; every dog owner knows there are things your dog does that your brain subconsciously understands, even if they cannot express in words what it is they are understanding, and every fighter pilot, tanker, or other military professional depending for their lives on a machine has a

[38] Crootof (2018, pp. 76–78).

[39] Wood (2023b, pp. 11–12).

sort of understanding for that machine, one bred not from textbooks and explanations but from sitting inside the thing and simply gaining an understanding.

Finally, there is the added problem that if XAI is achieved for some (set of) systems, there is a risk that this may perversely lead to *less responsible* deployments of AI systems. This is because overemphasis on explainability may lead XAI to be seen as a sort of "check box" for permissible use of AI systems. Yet, as argued above, it is possible for systems to be explainable in unhelpful ways, and it is possible that individuals better able to understand explanations may be less competent in actually predicting an autonomous agent's actions in dynamic environments. Thus, that AWS or military AI-systems are explainable in principle or practice may not imply that operators and handlers can understand the explanations or make reliable predictions based on them. The real efforts need to be in trust and teaming, not in technical accomplishments, and failure to do so can lead to disastrous consequences. As an example, consider the downing of Iran Air Flight 655, one of the deadliest military mistakes related to failures of human–machine teaming.

The crew of the USS *Vincennes*, a missile cruiser outfitted with a state-of-the-art Aegis combat system, misidentified a civilian airliner as an Iranian F-14, and due to "overconfidence in the abilities of the system, coupled with a poor human–machine interface", proceeded to engage and down the aircraft, killing the 290 civilians aboard.[40] While the systems on board the *Vincennes* were likely practically opaque at that time, there was nothing that would plausibly make them in principle opaque. More than this, it is certainly feasible that they could be made transparent and explainable using the current methods of XAI. But this is besides the point. The downing of Iran Air Flight 655 was not caused by opaque systems or a lack of understanding about the processes built into the Aegis combat system. It was the result of a series of failures in human–machine teaming and in cooperation between various combat units, and simply facilitating better communication between these groups would have allowed one to avoid the incident. Again, the core problem during deployments is not whether a system is explainable, but rather whether the system, explainable or not, is well-integrated into reliable human–machine teams which exhibit reasonable levels of trust and have individuals who know when and when not to rely on the system.

As a final word in this section, I again will stress that XAI does have value. That value is just not on the battlefield, but rather in design and troubleshooting labs. The upshot of this is thus not that we should abandon XAI, but rather that we should be cognizant of the limits of its benefits. If we are not, we may be blinded by an over-hyped research program and fail to recognize the extreme importance of other values (like human–machine teaming).[41] Moreover, we may find ourselves with an "ethical check box" which allows systems to be deployed to battle even when they have no one who can responsibly handle them or reliably predict their actions.

## Conclusion

XAI does have value in the military domain. By making opaque systems explainable, artificially intelligent systems may be more quickly, effectively, and reliably trained, designers may more rapidly remove processes that might lead to error or novel unwanted behavior, and the presence of understandable explanations can greatly streamline troubleshooting efforts when systems do act in unwanted ways. However, it is critical that we also understand that there are limitations to the good that XAI can provide. More than this, we must pay heed to the fact that implementing XAI in certain contexts has the potential to lead to mistakes as well, and that it may undermine the perceived need for highly trained handlers of AI systems who are intimately familiar with the capabilities, limitations, and quirks of reasoning in these systems. XAI can thus help us to more safely, reliably, and responsibly develop and maintain AI systems in the military domain (pre- and post-deployment value), but an uncritical implementation of these approaches across the board brings significant risks as well.

In closing, it is worth stressing that the overall intent of this article has not been to argue that XAI is a good or bad thing in the military, but rather to highlight that its value will be context-dependent and vary depending on who is engaging with the explanations provided. For engineers, designers, and troubleshooters, explanations are almost certain to be beneficial and ought to be incorporated to the greatest extent possible. However, XAI will do little to improve (or detract from) AI systems deployed without contemporaneous human oversight, and for those with a human in- or on-the-loop, XAI may create obstacles to responsible deployment of complex AI systems. Most importantly though, responsible deployments will require, first and foremost, strong human–machine teams where the human acts as a "handler" of the AI and not a "user" or "operator". In developing such teamings, looking to the analogy of human-animal teams in war provides a useful departure point and can inform us of the sorts of risks and pitfalls that are likely to arise if we treat potentially opaque AI systems as mere artifacts which

[40] Galliott (2020, p. 163). See also Sagan (1991, pp. 97–101) and Rogers et al. (1992).

[41] See Shneiderman (2022) for extensive elaboration of a similar general point.

can be easily understood and predicted provided one has an explanation of its reasoning processes.

**Acknowledgements** I must thank both Steven Umbrello and my wife Anna, both of whom listened to my first idea for this paper and encouraged me to pursue it. I am also deeply indebted to Maciek Zając, who was serving as a co-author for a time. We decided each of our papers would be better off standing alone, with more room to develop their respective ideas, but he helped this come to fruition in significant ways before moving onto his own manuscript, and for that I am very grateful. And finally, I must thank my dog Buddy, who somehow provided me with some of the most insightful thoughts of this article, and who is also a very good boy!

**Funding** A portion of this work was supported by the Grantová agentura České republiky (Czech Science Foundation), under Grant # 24-12638I.

**Data availability** Not applicable

**Code availability** Not applicable

## Declarations

**Conflict of interest** The author has no relevant financial or non-financial interests to disclose.

**Ethical approval** Not applicable

**Consent to participate** Not applicable

**Consent for publication** Not applicable

## References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival, 59*(5), 117–142.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Baker, D. (2022). *Should we ban killer robots? Political theory today*. Polity.

Blanchard, A., & Taddeo, M. (2022). Predictability, distinction & due care in the use of lethal autonomous weapon systems. *SSRN Electronic Journal*.

Bode, I., & Watts, T. F. (2021). Meaning-less human control: Lessons from air defence systems on meaningful human control for the debate on AWS. Technical report.

Boothby, W. H. (2016). *Weapons and the law of armed conflict* (2nd ed.). Oxford University Press.

Boulanin, V., Davison, N., Goussac, N., & Carlsson, M. P. (2020). Limits on autonomy in weapon systems: Identifying practical elements of human control. Technical report, International Committee of the Red Cross and Stockholm International Peace Reseach Institute.

Buchanan, J., & Kock, N. (2001). Information overload: A decision making perspective. In *Multiple criteria decision making in the new millennium: Proceedings of the Fifteenth International conference on multiple criteria decision making (MCDM)* Ankara, Turkey, July 10–14, 2000 (pp. 49–58). Springer.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. *Information Processing & Management, 60*(1), 1–16.

Caron, J.-F. (2020). Defining semi-autonomous, automated and autonomous weapon systems in order to understand their ethical challenges. *Digital War, 1*(1–3), 173–177.

Crootof, R. (2018). Autonomous weapon systems and the limits of analogy. *Harvard National Security Journal, 9*, 51–83.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey.

DeMay, C. R., White, E. L., Dunham, W. D., & Pino, J. A. (2022). Alphadogfight trials: Bringing autonomy to air combat. *Johns Hopkins APL Technical Digest, 36*(2), 154–163.

Ding, J., & Dafoe, A. (2023). Engines of power: Electricity, AI, and general-purpose, military transformations. *European Journal of International Security, 8*(3), 377–394.

Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy, 10*(3), 343–348.

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2021). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 19*(2), 133–144.

Flemisch, F. O., Adams, C. A., Conway, S. R., Goodrich, K. H., Palmer, M. T., & Schutte, P. C. (2003). The h-metaphor as a guideline for vehicle automation and interaction. Technical report.

Galliott, J. (2020). No hands or many hands? Deproblematizing the case for lethal autonomous weapons systems. In S. C. Roach & A. E. Eckert (Eds.), *Moral responsibility in twenty-first-century warfare: Just war theory and the ethical challenges of autonomous weapons systems* (pp. 155–180). State University of New York Press.

Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics, 4*(37), 1–2.

Gunning, D., Vorm, E., Wang, Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 1–12.

Haugh, B. A., Sparrow, D. A., & Tate, D. M. (2018). *The status of test, evaluation, verification, and validation (TEV &V) of autonomous systems*. Institute for Defense Analysis: Technical report.

Heller, K. J. (2023). The concept of "the human'' in the critique of autonomous weapons. *Harvard National Security Journal, 15*(1), 1–76.

Horowitz, M. C. (2020). AI and the diffusion of global power. *Modern Conflict and Artificial Intelligence*, 32.

Human Rights Watch. (2016). *Killer robots and the concept of meaningful human control*. Human Rights Watch: Technical report.

Hunter, C., & Bowen, B. E. (2023). We'll never have a model of an AI major-general: Artificial intelligence, command decisions, and kitsch visions of war. *Journal of Strategic Studies, 47*, 1–31.

International Committee of the Red Cross. (2014). *Autonomous weapons systems: Technical, military, legal and humanitarian aspects*. International Committee of the Red Cross: Technical report.

International Committee of the Red Cross. (2021). *ICRC position and background paper on autonomous weapons systems*. International Committee of the Red Cross: Technical report.

International Committee of the Red Cross. (2021). *ICRC position on autonomous weapons systems*. International Committee of the Red Cross: Technical report.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 1–24.

Layton, P. (2023). The age of war-fighting robots is upon us. *The Straits Times*. Retrieved September 4, 2023, from https://www.straitstimes.com/opinion/the-age-of-war-fighting-robots-is-upon-us

McCarthy, J. (1988). Mathematical logic in artificial intelligence. *Daedalus*, 297–311.

McFarland, T., & Assaad, Z. (2023). Legal reviews of in situ learning in autonomous weapons. *Ethics and Information Technology, 25*(1), 1–10.

Mecacci, G., & Santoni de Sio, F. (2019). Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics and Information Technology, 22*(2), 103–115.

Michel, H. (2020). *The black box unlocked: Predictability and understandability in military AI*. United Nations Institute for Disarmament Research (UNIDIR): Technical report.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Minsky, M. (1985). *The society of mind*. Simon & Schuster.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279–288). ACM.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Nowrot, K. (2015). Animals at war: The status of "animal soldiers'' under international humanitarian law. *Historical Social Research, 40*, 128–150.

Pacholska, M. (2024). Autonomous weapons. In B. Brożek, O. Kanevskaia, & P. Pałka (Eds.), *Research handbook on law and technology* (pp. 392–407). Edward Elgar Publishing.

Peters, U. (2022). Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque. *AI and Ethics, 3*(3), 1–12.

Phillips-Wren, G., & Adya, M. (2020). Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems, 29*, 213–225.

Roff, H. M., & Danks, D. (2018). "Trust but verify'': The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics, 17*(1), 2–20.

Rogers, W. C., Rogers, S. L., & Gregston, G. (1992). *Storm center: The USS Vincennes and Iran air flight 655: A personal account of tragedy and terrorism*. Naval Institute Press.

Ross, A. (2022). AI and the expert; a blueprint for the ethical use of opaque AI. *AI & Society*. https://doi.org/10.1007/s00146-022-01564-2

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Sagan, S. D. (1991). Rules of engagement. *Security Studies, 1*(1), 78–108.

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 1–14.

Scharre, P. (2023). *Four battlegrounds*. W. W. Norton & Company.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 2239–2250).

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics, 28*(5), 1–22.

US Department of Defense. (2023). DoD Directive 3000.09. Technical report, United States Department of Defense.

Verbruggen, M. (2022). No, not that verification: Challenges posed by testing, evaluation, validation and verification of artificial intelligence in weapon systems. In T. Reinhold & N. Schörnig (Eds.), *Armament, arms control and artificial intelligence, Studies in peace and security* (pp. 175–192). Springer.

Wang, P. (1995). *Non-axiomatic reason system: Exploring the essence of intelligence*. PhD thesis, Indiana University.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1–37.

Williams, A. P. (2015). Defining autonomy in systems: Challenges and solutions. In A. P. Williams & P. D. Scharre (Eds.), *Autonomous systems: Issues for defense policymakers* (pp. 27–62). NATO Communications and Information Agency.

Wood, N. G. (2023a). Autonomous weapon systems: A clarification. *Journal of Military Ethics, 22*, 1–15.

Wood, N. G. (2023b). Autonomous weapon systems and responsibility gaps: A taxonomy. *Ethics and Information Technology, 25*(1), 1–14.

Wood, N. G. (2023c). Rise of the machines or just a routine test? *War on the Rocks*. https://warontherocks.com/2023/06/rise-of-the-machines-or-just-a-routine-test/

Zając, M. Is LOAC compliance possible for AWS running unexplainable software? (Unpublished manuscript)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.