AI and Ethics (2024) 4:229–245
https://doi.org/10.1007/s43681-023-00261-0

ORIGINAL RESEARCH

Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability

Edward Hunter Christie[1] · Amy Ertan[2,3] · Laurynas Adomaitis[4,5] · Matthias Klaus[6,7]

Received: 4 July 2022 / Accepted: 6 January 2023 / Published online: 21 February 2023

Abstract
We explore existing political commitments by states regarding the development and use of lethal autonomous weapon systems. We carry out two background reviewing efforts, the first addressing ethical and legal framings and proposals from recent academic literature, the second addressing recent formal policy principles as endorsed by states, with a focus on the principles adopted by the United States Department of Defense and the North Atlantic Treaty Organization. We then develop two conceptual case studies. The first addresses the interrelated principles of explainability and traceability, leading to proposals for acceptable scope limitations to these principles. The second considers the topic of deception in warfare and how it may be viewed in the context of ethical principles for lethal autonomous weapon systems.

Keywords Artificial intelligence · Defence · LAWS · Ethical principles · Explainability · Deception

1 Introduction

There is a growing body of literature devoted to the ethical and legal aspects of the use of artificial intelligence (AI) for military purposes, with a dominant focus on the emergence of lethal autonomous weapon systems (LAWS) [1–3]. Reviewing the diverse field of disciplines contributing to the literature of military AI, Taddeo and Blanchard [4] found that depending on an organisation's approach towards LAWS, their definitions focus on different ethical and legal challenges. Distilling various definitions, they suggest defining LAWS as machines able to adapt to their environment and change between observation and engagement stances flexibly to identify, select and kinetically attack their targets without human intervention. It is this characteristic which distinguishes an autonomous weapon from an automated one, which has computational processes speeding up some

✉ Matthias Klaus
mk2124@cam.ac.uk; matthias.klaus@capgemini.com

Edward Hunter Christie
edward.hunter.christie@gmail.com

Amy Ertan
amyertan@pm.me

Laurynas Adomaitis
laurynas.adomaitis@cea.fr; laurynas.adomaitis@nordsec.com

[1] AI Policy Consulting, Brussels, Belgium

[2] Information Security Group, Royal Holloway, University of London, Egham TW20 0EX, England, UK

[3] Emerging Security Challenges Division, North Atlantic Treaty Organization, Brussels, Belgium

[4] CEA-Saclay/Larsim, 91191 Gif-Sur-Yvette, France

[5] NordVPN S. A., Fred. Roeskestraat 115, 1076EE Amsterdam, The Netherlands

[6] Leverhulme Center for the Future of Intelligence, University of Cambridge, Cambridge CB2 1SB, England, UK

[7] BTS Public Sector and Defense, Capgemini, Potsdamer Platz 5, 10785 Berlin, Germany

of its functions, but is only able to achieve narrow preset goals based on detailed and deterministic prior programming. As of 2022, the technological evolution is still ongoing from older, exclusively pre-programmed systems for narrow tasks towards more versatile systems that include greater adaptability, thanks to AI, for some of their functions (e.g. image recognition, voice recognition). In this article, we will adopt the definition of LAWS offered in Taddeo and Blanchard [4], while noting that such systems do not exist at this time (e.g. an unmanned aerial combat vehicle capable of carrying out an entire combat operation without human assistance). We further stress our understanding of lethality in the context of LAWS as referring to systems intended for military combat, namely including deliberate strikes on human combatants and on manned military platforms and vehicles.

One key debate within this literature deals with the question of whether international humanitarian law (IHL) in its current form, as the most relevant body of law concerning warfare, sufficiently covers the challenges of LAWS [5, 6]. To this end, this article first documents the approaches of relevant organisations and nations towards military AI and highlights how they interrelate with and refer to IHL.

In recent years, several states in the international system have made official political commitments on the ethical use of AI in defence. In particular, it is worth noting the principles of ethical AI adopted by the United States Department of Defense (DoD) in February 2020 [7], which were based on a report by the Defense Innovation Board [8, 9]. A further major development was the adoption by all 30 governments of the North Atlantic Treaty Organization (NATO) of a similar set of principles, as part of the adoption of NATO’s first-ever artificial intelligence strategy [10, 11]. Both the US principles and the NATO principles apply to all military applications of AI, including but not limited to LAWS and will be our focus in this regard.

Following the literature review and outline of relevant organisations and national positions, we will investigate one NATO principle in particular, namely *Explainability and Traceability*, and explore interrelated questions that pertain to it relating to transparency, security, and intended versus unintended unpredictability and deception. We note that they are technically linked with certain types of AI, notably machine learning. More specifically, a machine learning algorithm whose parameters are fully observable by all parties would be explainable and traceable, and thus be compliant with the principle. However, its behaviour will also be—in principle—predictable for all parties, meaning it would violate security needs and offer opportunities to the adversary to predict its actions and gain an unwanted military advantage. For purposes of military effectiveness, a LAWS should be unpredictable for the adversary, and possibly for the party that operates it, within certain limits. The highest degree of military effectiveness would be achieved if the LAWS could successfully deceive the adversary, while remaining within the bounds of IHL and while also posing no danger to the side that operates it. As these considerations illustrate, it is relevant to explore in more detail the boundaries of these related concepts.

2 Academic research on ethical frameworks for AI in defence

We make the working assumption that the emergence of LAWS will be technically possible and, furthermore, that military necessity and economic efficiency will make the development and employment of certain kinds of LAWS inevitable [12, 13]. Our chosen focus is on the regulation of LAWS in the context of warfare, where the behaviour of states is regulated by IHL. States pursue their IHL obligations through relevant national guidance documents that place outer limits on the rules of engagement they adopt ahead of the use of military force. In the absence of new normative frameworks, only IHL and national rules of engagement will constrain how LAWS are deployed and used [14, 15].

These core principles apply regardless of the means of warfare, and therefore also apply to the use of LAWS, as confirmed by states participating in the UN CCW GGE on LAWS. A challenge for the application of these core principles to the case of LAWS is that IHL is targeted at human action and presupposes human intent for an action to become judgeable, something AI systems lack [16]. For clarity, all existing law is predicated on human responsibility and accountability, and there is no serious debate at this time about holding an AI system or a LAWS legally accountable for anything. However, as LAWS may select their own courses of action within designated parameters, they may display complex and unintended behaviours in the pursuit of military objectives which might threaten compliance with IHL principles, such as proportionality.

Even before the World Wars of the twentieth century and the concomitant introduction of new ways of industrialised killing, early forms of IHL like the St. Petersburg declaration attempted to regulate warfare with regard to proportionality [17]. Later on, more specific documents followed suit by completely banning the use of certain weapon systems. This includes biological and chemical weapons, as they could not be employed in proportional and distinct manner, among other reasons also involving humanity.[1] Similarly, ‘dumb’ weapons like cluster ammunition employed during

[1] See the 1972 Convention for example: https://www.un.org/disarmament/biological-weapons.

nightly air bombings came under scrutiny [18], which contributed to the development of 'smarter' alternatives like GPS or laser precision-guided munitions. These weapons can enable targeting with pinpoint accuracy, which contributes to their higher IHL compliance. The precision of these systems, together with qualities such as adjustable fuse settings, allows users to achieve their military goals with minimal force, thus reducing the exposure of the target's surroundings to military action. Nevertheless, operator errors or mechanical failures can and have resulted in many publicised incidents where supposedly 'smart' bombs caused collateral damage, i.e. damage to unintended, often illegitimate targets. LAWS have the potential of constituting another step forward in ensuring greater accuracy and less collateral damage, meaning greater compliance with IHL, most notably with the principles of distinction and proportionality. LAWS could employ precision munitions autonomously and process information faster than any human operator [13, 19], but their opaqueness and unpredictability in unforeseen situations pose different kinds of risk [20, 21].

Many authors focus on the implications of LAWS with respect to the IHL principles of proportionality and distinction [6, 19, 22, 23]. Under IHL, human commanders are expected to demonstrate the reasonableness of their attack decisions to explain and justify their conduct. The black box problem, that is, the opaqueness of modern machine learning algorithms, which prevents any human from understanding their decision-making process, poses severe ethical problems in this regard, as one cannot predict how an AI would pursue its task [20]. The lack of 'intent' of AI systems contributes to this dilemma, as it cannot be proven that an opaque and indifferent AI system would act in good faith, as required by the proportionality principle. Some authors [13, 24, 25] suggest that narrowing the freedom to manoeuvre could help overcome these challenges. It could involve controls on the weapon system's parameters to restrict targeting and tasks to only certain target sets like navy vessels, on the environment of use like the open sea, and requiring human–machine integration that ensures human control via in-the-loop or on-the-loop oversight. These concepts describe the level of integration of the human operator into the functioning of the weapon system. In-the-loop means that the operator has to confirm any potential attack, while on-the-loop allows the system to attack autonomously with the human supervising in the role of fail-safe. Finally, off-the-loop (or out-of-the-loop) describes a system operating fully autonomously without a human supervising its actions. The ICRC also recommends developing practical human control measures, coupled with internationally agreed constraints on LAWS based on IHL, and clarification where new rules might need to be developed [26]. Müller [21] echoes this by pointing out that the distinction between military and civilian vessels could be easier than between soldiers and non-combatants. In sum, LAWS are a challenge for responsibility and accountability under IHL, as they could replace humans in both action and planning. Even if a human is in-the-loop, the amount of data and necessary speed of decision-making severely limit human oversight [6, 13, 21]. The question of responsibility becomes even more pressing when there are no humans in- or on-the-loop. Arkin [14, 22] argues that the responsibility gap can be bridged by assigning responsibility advisors along the whole process, ranging from design to employment.

Based on these discussions, the ICRC and SIPRI argue that a key necessity for LAWS in war is human control [26, 27], which is echoed by other authors [19, 28, 29]. Related literature highlights the importance of "meaningful human control", or MHC, of AI systems in military operating environments [30, 31]. MHC consists of the role of the human operator as a fail-safe to ensure ongoing compliance with IHL, to ensure ongoing accountability for possible breaches of IHL, and to ensure human moral agency remains involved with any actions taken by LAWS [32]. Roughly, there seems to be two schools of thought regarding the demands IHL has towards human control of LAWS. One school of thought argues that technological development can overcome the existing challenges to MHC, as improvements in the field of AI will allow for systems to adhere ever more closely with IHL principles while retaining the advantages of faster decision-making speed and objective decision-making processes [14, 20, 24, 33]. Assuming further advances in AI, one could imagine machines being better at complying with IHL than humans, although this is as yet a distant prospect [34]. The other school of thought argues that IHL should be interpreted as requiring inherent limits to the autonomy of LAWS, because IHL principles can only be fully met based on contextual and ethical judgements by humans [6, 12, 26]. Also, LAWS should not be employed unless MHC can be enforced to close the gap of responsibility [35].

While human–machine integration is proposed as an efficient way to mitigate many existing concerns about MHC, human operators could suffer from various biases [21]. While automation bias may result from too much trust in sophisticated AI [23], the reverse can be true due to the inherent opaqueness of these systems [20]. Sullins [36] raises the notion that AI will need to learn and understand deception in warfare to compete with human counterparts who use these tactics routinely. This is a dilemma in itself, as it presupposes humans building machines with traits we consider unethical. Focusing on how AI could be used to facilitate unacceptable military deception, Chelioudakis [37] found IHL was flexible enough to account for, and remain unchallenged by, deceptive AI machines available at the time. We will explore the topic of deception in more depth later in this paper. Arkin [22] also agrees that IHL

Table 1 Core principles of IHL

Core principle	Definition
Proportionality	Incidental damage to civilian structures and injury or death of civilians must not be excessive with regard to concrete and direct military advantage
Precaution	Those who plan or decide upon an attack shall take all feasible precautions. All parties must verify their targets and warn the civilian population before attacking, unless circumstances do not permit. Further restrictions on attack timings, location and other characteristics can become necessary
Distinction	Distinguish at all times between civilians and civilian objects—and combatants and military objectives, only the latter two may be targeted. Prohibits indiscriminate attacks
Military necessity	Allows for measures to achieve a legitimate military goal which are not prohibited by IHL. Needs to be balanced against humanitarian concerns to prevent unnecessary suffering
Humanity	In cases not covered by IHL, persons affected by armed conflicts will still be protected by the laws of humanity and public conscience ("Martens Clause")

Source: ICRC (See https://casebook.icrc.org/glossary/fundamental-principles-ihl) and Geneva Protocols (See https://www.icrc.org/en/doc/war-and-law/treaties-customary-law/geneva-conventions/overview-geneva-conventions.htm)

is sufficient to constrain LAWS, arguing that while LAWS could act unethically, they would still perform more ethically than human soldiers. However, lacking experience with the (disruptive) application of AWS [14] makes it unlikely that LAWS could comply with IHL at the current stage. This is in contrast to Anderson, Reisner, and Waxman [38], who propose adapting IHL to account for greater autonomy on the battlefield.

One should also make note of interpretations of how the IHL principle of humanity should apply in the case of LAWS. For some authors, it does not seem possible to respect human dignity in the case of a LAWS which would take a kill decision based on a black box algorithm [39]. A counterpoint to this view could be made from counterfactual analysis. A putative high-performance LAWS with a track record superior to humans in respecting the IHL principles of distinction and proportionality could be thought of as being in better compliance with the principle of humanity from a utilitarian perspective, e.g. less collateral damage, faster and more effective fire against legitimate enemy targets, and self-restraint for reasons of proportionality.

Overall, the need for various kinds of human control during both the design and use of LAWS is a clear insight from existing literature. Assuming that LAWS would use AI based on existing classes of machine learning algorithms, the latter's inherent lack of contextual understanding and black box characteristics leave very few scenarios in which LAWS could operate fully on their own. We recognise that this situation could result in a military disadvantage for a state that takes IHL compliance seriously, but that may face an opponent who develops or uses LAWS with a low level of regard for IHL.

As reviewed above, available literature has evolved towards a general acceptance that the existing five core principles of IHL shown in Table 1 are the most relevant and suitable set of objectives that future LAWS should be in compliance with. The heart of the debate among experts revolves around additional guidance or rules that would be specific to LAWS and that would further specify how to ensure that LAWS are developed and used such that IHL compliance is both enabled and facilitated. We will address the positions of selected governments and international organisations in the next section.

Autonomous systems are intended to be able to navigate complex and uncertain environments and to determine, by themselves, successful courses of action to fulfil their missions. The courses of action a future LAWS will choose under real battlefield conditions will have elements of unpredictability for any human observing it, including its human commander—much like a soldier should be able to improvise on the spot to pursue mission objectives, sometimes in ways that could surprise its commander. The AI software that will enable a future LAWS to complete missions successfully is likely to be highly complex and not amenable to simple technical standards that could guarantee compliance with IHL while allowing the AI software to have the required degree of versatility. There is no guarantee at this time that future, highly advanced AI will be developed to have such a high degree of semantic and contextual understanding that software could simply be instructed to "understand IHL" and comply with it at all times. Assuming current or near-term foreseeable technologies, the design of a LAWS will require various forms of MHC, and the use of a LAWS will require human operators in- or on-the-loop in all scenarios in which autonomous operation presents a significant risk to IHL compliance. As a result, and as we will see in the next section, commitments and practical work currently being undertaken by several states go much beyond binary questions about the applicability of IHL, that question being already settled positively, and towards additional principles specific to LAWS, or to military applications of AI more broadly, as well as towards processes to operationalise such principles.

3 Official positions of national governments

The relevant intergovernmental forum for the negotiation of norms that could place limits on the development or use of LAWS is the group of governmental experts on lethal autonomous weapon systems (GGE LAWS) that operates in the framework of the Convention on Certain Conventional Weapons (CCW). The CCW itself is an international treaty and a set of additional protocols that set out norms and limitations on certain kinds of conventional weapons. As of 2022, 126 states are parties to the CCW,[2] including the five permanent members of the Security Council (USA, Russia, China, UK, France), most other significant military powers, e.g. India, Pakistan, Israel, Saudi Arabia, all members of the G20 group of the world's largest economies (except Indonesia), all members of NATO, all members of the European Union, and all Latin American countries. Overwhelmingly, those states that are not party to the CCW are nations in Africa, the Caribbean, the Middle East, and Southeast Asia. From a hard security perspective, the most notable non-parties are Iran and North Korea. The GGE LAWS started operating in 2014, with a majority of state parties participating. A pivotal development was the adoption of 11 guiding principles that were proposed in the conclusions of a 2018 report of the GGE LAWS [58] and adopted by consensus by participating state parties at the group's November 2019 meeting [59].

The first 2 of the 11 principles are the most foundational. The first affirms that IHL "continues to apply fully to all weapons systems, including the potential development and use of" LAWS. The second affirms that "human responsibility (…) must be retained since accountability cannot be transferred to machines". The other principles derive from the first two principles and address, in broad terms, the need for measures addressing both the development and the use of LAWS, including the desirability of risk assessment, risk mitigation, and security measures (including cybersecurity). The topic of human control is addressed in the third principle under a heading of "human–machine interaction" which stresses IHL compliance but does not spell out specific limits based, e.g. on certain functions being necessarily subject to in-the-loop versus on-the-loop human supervision. Overall, the 11 principles do not constitute new legally binding obligations but a consensus among state parties regarding what they each commit to uphold in their national practices. The principles also do not open up any prospects for an international monitoring or verification regime for LAWS. As agreements among state parties are by consensus, the maximum extent of agreed limitations on LAWS are the lowest common denominator between national positions. As of 2019, therefore, the trajectory for an international agreement seemed set on merely affirming IHL, with flexibly worded political commitments on human control. There was no sign, then, of any possibility of having the five permanent members of the UN Security Council, in particular, agreeing to a consensus on wide-ranging prohibitions on either the use or the development of certain kinds of LAWS. We attribute that pattern to a combination of two factors: rival major military powers, being in a state of mutual distrust, wish to retain flexibility to explore the military advantages they may each achieve—bearing in mind that LAWS have the potential, through greater accuracy and speed, of being superior to non-autonomous equivalents both in terms of their effects on opposing forces as well as in terms of the lesser danger they may pose to one's own forces (in great contrast to chemical or biological weapons, which do not have such characteristics and which are prohibited).

By 2022, however, certain national positions had evolved. Two groups of countries, the USA, the UK, Korea, Japan, and Australia, on the one hand, and Argentina, Costa Rica, Guatemala, Kazakhstan, Nigeria, Panama, Philippines, Sierra Leone, State of Palestine, and Uruguay, on the other, each submitted a joint paper [78, 79] to the GGE LAWS that had in common proposals to prohibit four potential types of LAWS. Using the wordings from the first paper, these are:

(1) LAWS "of a nature to cause superfluous injury or unnecessary suffering [or] if it is inherently indiscriminate, or if it is otherwise incapable of being used in accordance with international humanitarian law".
(2) LAWS "designed to be used to conduct attacks against the civilian population, including attacks to terrorize the civilian population".
(3) LAWS "designed to cause incidental loss of civilian life, injury to civilians, and damage to civilian objects that would invariably be excessive in relation to the concrete and direct military advantage expected to be gained".
(4) LAWS in which the autonomous functions are designed "to be used to conduct attacks that would not be the responsibility of the human command under which the weapon system would be used".

It can be argued that the four types defined above necessarily derive from applicable IHL and from the second guiding principle adopted in 2019, namely that accountability cannot be transferred to a machine. Nevertheless, explicit prohibitions provide for greater legal clarity and have a commitment value between states as well as towards populations and civil society.

[2] For the list of state parties, see: https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/high-contracting-parties-and-signatories-ccw/

On the other hand, the aforementioned paper by the USA and some of its allies effectively allows for the potential use of a LAWS that could autonomously engage military targets in accordance with IHL, that is, without necessarily having a human in-the-loop.

While other national submissions made to the GGE LAWS in 2022 also contained important elements, the overlap between the two joint papers described above represents the strongest potential to date for agreed prohibitions on the use of certain kinds of LAWS.

In parallel with the potential future development of an intergovernmental agreement, which could take the form of a new protocol under the CCW, states have also been working on 'soft law', such as national guidelines and principles that are not legally binding, at the national level as well as at the NATO level for those states that are members of NATO. Soft law approaches are important in that they can provide more detailed guardrails to structure national activities than what states may be comfortable agreeing to in a legally binding convention or treaty.

The USA was the first state actor to release a defence-specific AI strategy, with the DoD publishing the strategy's executive summary publicly in February 2019 [8, 9]. This was followed in 2020 by the DoD adoption of five 'AI Principles', which outline that deployed AI must be: responsible, equitable, traceable, reliable and governable.[3] These principles apply "to all DoD AI capabilities, of any scale, including AI-enabled autonomous systems" as confirmed via a Biden administration memorandum [40]. While the memorandum confirmed the focal point for the development of responsible AI infrastructure would be the Joint Artificial Intelligence Center (JAIC), from 1 February 2022, the office of the newly established Chief Data and AI Officer has taken on these responsibilities. While there is no all-encompassing national guidance regarding LAWS, the US Directive 3000.09 on Autonomy in Weapons Systems [42] sets out a framework for using LAWS in ways that are consistent with IHL and any applicable treaties and rules of engagement (RoE). The Directive states that LAWS must be designed so commanders and operators are able "to exercise appropriate levels of human judgement in the use of force". While the Directive is due to be updated (as of late 2022), it is not expected to change significantly in form [83]. Consistently, the US view to date has been to favour a dispersed model of human judgement, whereby humans do not necessarily have to be in charge at the specific moment of engagement, but at crucial moments throughout the process [41, 43]. Senior US officials have stated that while the USA does not currently have fully autonomous LAWS, they may develop such capabilities if US competitors choose to do so [84].

In June 2022, the UK published its Defence Artificial Intelligence Strategy [44] alongside corresponding policy paper which set out five ethical principles for defence: human-centricity; responsibility; understanding; bias and harm mitigation; and reliability [45]. Both documents showcase a range of commitments concerning the development of responsible AI tools while stressing the CCW as the primary avenue for discussions on LAWS and IHL. The Strategy reinforces the position set out in the UK Ministry of Defence Unmanned Aircraft Systems doctrine that autonomous or remote-controlled systems must be operated in accordance with existing domestic and international legal frameworks, including IHL [44, 46]. The policy paper also reaffirms, word for word, the statement set out in this doctrine that "the UK does not possess fully autonomous weapon systems and has no intention of developing them" [45, 46]. Article 36 of the 1977 Additional Protocol I to the Geneva Conventions (hereafter: Article 36) is interpreted broadly by the UK, to include considerations of how weapons are used and the conduct of warfare, in which "weapon reviews may record limitations on the use of a weapon or method of or means of warfare in order to ensure compliance" [47].

France hosts a Defence Ethics Committee within its Ministry for Defense which in 2021 set out their view on the integration of autonomy into LAWS in a committee report [48]. In the report, the committee argued that while fully autonomous weapons are ethically unacceptable, reaffirming French national position since 2013, *partially* autonomous weapons systems may be ethically acceptable subject to defined '5C' conditions: command, risk control, compliance, competence, and confidence [48, 49]. Alongside such conditions, the committee recommends that a complete legal review is conducted wherever decision-making autonomy is developed in a lethal weapons system [48]. French definitions equate the UN definition of LAWS to fully autonomous systems and France has concisely rejected incorporating this form LAWS into military operations [48, 49].

China has published several documents relating to AI governance (not specifically for defence), including a set of principles with 'Chinese characteristics' like harmony [51]. In 2019, China established a National Ethics Committee on Science and Technology to supervise the regulation of AI in general. Commentators have pointed out that China's definition of LAWS is ambiguous, potentially allowing for machines that could not be deactivated or that could use force indiscriminately [53]. In a position paper on military AI regulation released in early 2022, China refers to the broader need to manage potential risks, but sets out no specific commitment or initiatives that suggest the development of national laws, rules or regulations for LAWS [85].

[3] For more detailed description of the Ethical Principles, see: https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

Russia has argued that IHL in its current form is sufficient for regulating LAWS [54] and has actively and consistently engaged in GGE discussions to oppose the establishment of legal binding instruments relating to LAWS [86]. Russia further contends that "excessive regulation can hamper the pace of development of EDTs, including AI" [55], and argues in favour of LAWS over human controlled machines due to their overall better performance, especially with regard to IHL compliance. Moreover, Russia commits itself to Article 36, dealing with responsibility and accountability, arguing that universalising the Article would suffice in contrast to specifically designed legal review measures [56]. This has been interpreted as an attempt to narrow international regulation with the goal of continuing domestic development unabated [57].

4 Soft law

NATO's first AI Strategy was agreed by Allied Defence Ministers in October 2021, with a public summary document outlining the Alliance's aim "to lead by example and encourage the development and use of AI in a responsible manner" [10]. The Strategy contains a set of "Principles of Responsible Use" as listed in Table 2. The principles were closely modelled on the DoD principles and other existing national principles, and apply to all kinds of AI applications that are intended for deployment (*ibid*).

Existing national soft law commitments and principles are rooted in considerations of existing IHL and of the need to always maintain human accountability, but they have the potential to go further and to be more granular than intergovernmental agreements at the UN. As of 2022, the US and NATO principles, which are highly similar, provide guidance for ongoing applied work at national level, which can be expected to take the form of more detailed national (and NATO) manuals, processes, and standards. We see these processes developing, for example, through the October 2022 releases of the NATO Autonomy Implementation Plan [81] and announcement regarding the creation of the NATO Data and Artificial Intelligence Review Board [82], both of which focus on operationalising NATO's Principles for Responsible Use as set out in Table 2. It is with such ongoing developments in mind that we wish to pose more detailed questions regarding one NATO principle in particular, namely that of explainability and traceability, owing to our estimation that it carries inherently more complex implications than the other NATO principles.

5 Interpretation and meaning: explainability and traceability

Both NATO and DoD frameworks include traceability among their principles of ethical and responsible use. NATO principles also mention explainability in addition to traceability. In civilian ethical guidelines for AI, traceability and explainability correspond to transparency, which is one of the most frequently mentioned principles [60]. However, transparency is an ambiguous term. Therefore, it will lend our further discussion to carefully distinguish and define traceability, explainability, and transparency.

Transparency is commonly understood as having and revealing information about internal processes of a public institution, a company, or other enterprise. This type of transparency is often considered a virtue that lends to fighting corruption, enacting accountability, and enhancing trust. The most commonly found definition of transparency relies on an enterprise's responsibility to make some information publicly available. It is usually formulated from the sender's (enterprise's) perspective without involving the responsibility to ensure the receiver (the public) is actually informed [61].

> Strict transparency: a process is transparent *if* information about the process is made publicly available.

We shall call this definition *strict* to distinguish its explicit part (making the information public) from its implicit parts (that the information exists and what information it is). Strict transparency simply captures the imperative to share information with the public or the stakeholders. It represents the necessary condition of transparency, for an enterprise that has information but does not make it publicly is not generally considered transparent.

However, in some AI applications, the information about the functioning of an algorithm can be unattainable. For example, black box systems are unexplainable. An enterprise could reveal the fact that a black or grey box algorithm is used, but they cannot explain how a decision process is carried out. Also, strict transparency is traditionally formulated from the sender's perspective, which presupposes that the receiver will find the information understandable. However, many AI applications can often provide only a quantitative reason why certain inputs and outputs are correlated. Such explanations do not translate into a semantic explanation for the stakeholders.

Two notions that address the issues above are explainability and traceability. Traceability means that certain outputs from an AI algorithm can be traced to certain inputs, as if going back in the decision chain. Traceability provides the ability to find a responsible input A for the eventual output B. From an ethical point of view, traceability

Table 2 NATO principles of responsible use of artificial intelligence in defence

Principle	Definition	Comparison with US DoD principles
Lawfulness	AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable	The preamble text to the DoD principles states that they will “build on the U.S. military’s existing ethics framework based on the U.S. Constitution, Title 10 of the U.S. Code, Law of War, existing international treaties and longstanding norms and values”
Responsibility and accountability	AI applications will be developed and used with appropriate levels of judgement and care; clear human responsibility shall apply to ensure accountability	Equivalent to the DoD’s “Responsible” principle. The linkage with the notion of accountability is made explicit. The term “human” is included for clarity; however, the DoD principle implies it through the use of the term “personnel”
Explainability and traceability	AI applications will be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. This includes verification, assessment and validation mechanisms at either a NATO and/or national level	Equivalent to the DoD’s “Traceable” principle. A clarification is added that relevant work may occur at the NATO level, or nationally, or in some combination of the two
Reliability	AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures	Equivalent to the DoD’s “Reliable” principle. Relevant work may occur at the NATO and/or national levels. The notion of certification is explicitly mentioned
Governability	AI applications will be developed and used according to their intended functions and will allow for: appropriate human–machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour	Equivalent to the DoD’s “Governable” principle. An explicit mention is added to human–machine interaction
Bias mitigation	Proactive steps will be taken to minimise any unintended bias in the development and use of AI applications and in data sets	Equivalent to the DoD’s “Equitable” principle. An explicit mention is added to the role of data sets

Source: [80]

can be significant in ascribing responsibility and predicting future behaviour (governability). However, knowing that A and B have been correlated does not eo ipso give an explanation of why they were correlated. Explainability refers to the ability to provide a semantic expression (as opposed to merely quantitative and operational) to why decision processes developed in a certain way. It is important to note that semantic expressions are not necessary for AI's functioning [62]. They need to be superadded to the technical characteristics.

> Traceability: a process is traceable if certain outputs can be correlated to certain inputs.

> Explainability: a process is explainable *if* (1) it is traceable and (2) the correlation can be given a semantic expression.

Strict transparency, traceability, and explainability are conceptually related, together implying a broader notion of full transparency, encompassing both the implicit premises about the attainability of information and its sharing. Traceability is necessary, but not sufficient for explainability. Explainability is necessary, but not sufficient for full transparency. Full transparency is implied by explainability and strict transparency.

Although civilian ethical guidelines focus on the disclosure of information, i.e. Strict Transparency, the defence sector uses classified information that is not generally released to the public, so strict transparency cannot be a requirement. Therefore, NATO and DoD principles do not include "transparency" among the principles for responsible use. It is replaced by traceability and explainability. Below, we will focus on the latter two.

The principal reason for wanting traceability or explainability in LAWS is the enabling function of these principles for the other areas of responsible use. If there were no traceability of inputs and outputs in the algorithm, one could not determine the actors in a decision chain, which makes accountability difficult to enforce [63]. Bias mitigation requires explainability to identify the bias groups and know that a decision was made because of the bias [64]. Traceability is the key to governability because the latter relies on predicting future outputs, but that means knowing that a certain input will produce a certain output, which is exemplified in CS8 and CS9 above.

The fact that traceability and explainability facilitate the operationality of other principles makes them a justified addition to the guidelines on LAWS. However, we shall pursue a critical analysis of the limits of their implementation. The main issue to consider is whether these principles should be held absolute and necessary for LAWS. We have already established the reasons for them being desirable but that does not mean they should be necessary or absolute.

Explainability is de facto not absolute in the state of the art for most deep learning algorithms. Many successful applications of deep learning do not require explanations, including military applications [65], and enforcing explainability constraints may decrease performance. There are scholars who claim that deep learning algorithms are inherently not explainable, thus ascribing semantic explanations to them is at best a plausible story that cannot be proven [51]. LAWS require AI applications in computer vision, robotics, and decision systems, which often depend on deep neural networks, and thus are not currently explainable.

If the requirement of explainability would be absolute, it would impede the use of LAWS. However, advanced AI systems can be much more efficient than deterministic models in real world applications, so whichever party is using unexplainable deep learning in LAWS can have a significant advantage. Therefore, an expectation for the absoluteness of explainability is not compatible with the use of LAWS and produces a military disadvantage.

In view of this, NATO principles rightly consider only an "appropriate" level of explainability and do not require absolute explainability, therefore implying that algorithms do not have to be absolutely explainable or traceable. They need to be explainable to a certain degree. The details of the particular level of explainability are ingrained in the validation procedures but are undefined in the public documentation. The text claims "AI applications will be appropriately understandable", which seems to imply that an algorithm would need to pass some benchmark, so traceability or explainability must be higher than zero for responsible use. If this is true, then NATO embraces the necessity but not absoluteness of explainability for responsible use.

However, the wording leaves room for discussion. "Appropriately understandable" may allow that for some algorithms the appropriate level of understandability is zero. We suggest this is a reasonable proviso to include in the guidelines. The necessity of explainability would rule out the use of black boxes, which can be detrimental to performance and responsible use.

In the case of LAWS, where lethal force is involved, there seems to be an intuitive need for explaining every step of the decision chain [66]. Traceability and explainability feel crucial to find the accountable decision makers and to explain why an accident has happened. A semantic explanation feels necessary for anthropocentric reasons. A term "semantic anthropocentrism" [67] has emerged in the literature to describe the need for human explanations from other intelligences, including AI, although semantic explanations often rely on borrowing concepts from human and animal cognitive science and their application to domains like LAWS can only be metaphorical. Contrary to the intuition, we propose three main arguments against the necessity of explainability and in favour of the use of black boxes in LAWS.

5.1 Accuracy argument

Firstly, there are significant use cases where black boxes are more accurate than xAI.[4] Scholars generally agree that “there is clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions” [68]. Semantic explanations cannot be given for the decisions of deep learning algorithms because of their complexity. The trade-off for the higher accuracy is very low or non-existent explainability. If the necessity of explainability was enforced in the NATO principles, this trade-off would be seen as unacceptable. However, in the use of LAWS, where the stakes often involve lethal consequences, the trade-off of explainability for accuracy seems reasonable. Accuracy in LAWS can translate into a successful military operation instead of an unsuccessful one or a defeated enemy instead of a civilian casualty. Thus, in trading off accuracy, one would be creating an increased danger of unintended lethal damage. The gain of explainability from this trade-off is not considered a sufficient justification because its benefits are conceptual and not material, i.e. explainability helps to understand why an error happened after the event. The potential negative effect of this trade-off (decreased accuracy) could have a compounding nature, where one less successful operation leads to an advantage for the adversary that can cause more lethal damage to the ally. Moreover, in scenarios where both adversaries use LAWS, the one with more accuracy will likely prevail which can result in a military loss. The trade-off of accuracy for explainability is problematic in LAWS, because the impact of accuracy metrics is material and lethal, and the advantages of explainability are only conceptual.

5.2 Secrecy argument

Secondly, the robustness of an AI system can rely on a degree of opacity that prevents malicious actors from successfully reverse engineering it. There are principles in cryptography, e.g. Kerckhoffs's principle, which require cryptographic systems to be safe even if all information about the system is public, and in fact the development of AI systems can heavily depend on open source material. However, military technology can require secrecy. Many of the robustness measures in AI safety only work well when the adversary does not know about the use of these measures and they fare extremely poorly after an analysis.[5] Since current innovation in AI robustness mostly takes place in public research, and the turnover of attacks and defences is very short, it is difficult to create cyber defences for LAWS that could not be revealed by open source intelligence [69]. Even if LAWS were equipped by completely classified robustness measures, military espionage and congeniality (the likelihood of finding a similar solution independently by an adversary) would remain a significant threat. Deep learning algorithms are not immune to adversarial attacks; however, they avoid producing explainability metrics that would involve analysing and visualising the architecture of LAWS. If this explainability information were to leak, it would lead to a significant threat to LAWS security. Avoiding producing this information in the first place can be a means of avoiding security by obscurity, i.e. relying on safeguarding technical information from adversary intelligence. At the same time, appropriate testing and benchmarking techniques need to be applied to ensure the accuracy and robustness of black boxes while treating them as such [70].

5.3 Trust argument

Lastly, contrary to common opinion, black boxes can elicit trust. It is commonly held that an explanation of a mechanism is necessary for trusting a system [71]. However, it is not always the case, and there is another source of trust—practical value. Black box algorithms have been used in high-risk applications, including military, healthcare, and criminal justice. These systems often function with human oversight but without being explainable. Practitioners in these fields (e.g. radiology) rely on them for expertise. If a system continually produces accurate predictions, it is bound to elicit trust. The emerging trust is expected to primarily appear in people who work directly with the systems. Researchers may remain sceptical for theoretical reasons, the public may remain critical because of ideological reasons but a functioning system will eventually elicit trust in the practitioners. This is observed in GGE conclusions, namely, “if the automated assessment has a very low rate of “false positives” […], the operational context corroborated the automated assessment, and the context involved combat operations, then it would seem to be reasonable to rely on the assessment [and] strike the target.”[6] Accuracy simply implies reasonable trust in an operational context.

Many researchers that take a critical stance towards transparency practises, even when it is to improve them, face the morally partial connotation of the notion of transparency. As Lord has put it, “transparency comes loaded with normative

[4] Some research in xAI claims that this trade-off is a myth, but clearly the myth is the universalised version of this claim that all xAI has this trade-off. The latter claim is not required for our argument.

[5] See https://www.robust-ml.org/defenses/.

[6] See https://geneva.usmission.gov/2021/08/05/u-s-statement-at-the-gge-on-laws-during-the-discussion-of-agenda-item-5d/

baggage. Like security, it is hard to be against transparency. Who is in favor of concealment or censorship?" [72]. However, NATO and DoD principles show how accountability, governability, and other ethical principles can be enabled without the use of Strict Transparency or Full Transparency. The terms seem almost superfluous for the ethical analysis in defence. We have argued that even the substituting principles of traceability and explainability should not be considered as absolute or necessary, but only as desirable. The framework for responsible use of LAWS should leave room for the use of black boxes, although a full set of validation techniques needs to support their deployment.

However, in addition to accuracy requirements, which seem to override the imperatives for explainability, there are links between explainability and governability that may present a different case. If LAWS are indeed autonomous, they will have to come up with winning strategies autonomously, i.e. without explicit human input. Winning strategies will be decided based on metrics, which in warfare is often understood as surviving and eliminating targets. The ethically challenging case is that of AI algorithms attaining their metrics by employing manipulation. For example, we know that chatbot algorithms competing against each other can learn "to deceive without any explicit human design, simply by trying to achieve their goals" [73]. Thus, there is reason to believe LAWS would learn deceptive strategies to defeat the adversary. However, deception strategies in warfare are specific and regulated under IHL. How can we make sure that these regulations will be observed by the winning deception strategies invented by LAWS? This task seems to require the understanding—and explanation—of how they come up and enact such strategies.

6 LAWS and military deception: a thought experiment

Military deception is as old as warfare [74]. In modern warfare, and in related IHL documentation, the term 'ruses of war' is used to encompass those acts of deception which are lawful, whereas the terms 'perfidy' and the adjectives 'perfidious' and 'treacherous' are used to describe acts of deception that are unlawful. Notably, Article 37(1) of the 1977 Additional Protocol I to the Geneva Conventions prohibits the killing, injuring or capturing of an adversary "by resort to perfidy" and defines perfidy as "acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence". On the other hand, Article 37(2) of the Protocol states that "ruses of war are not prohibited. Such ruses are acts which are intended to mislead an adversary (…) but which infringe no rule of international law (…) and which are not perfidious (…) [such as] the use of camouflage, decoys, mock operations and misinformation". More granular lists of examples of ruses of war can be found in guidance manuals produced by individual nations for their armed forces. Such lists are highly similar between countries (ICRC, n.d.). For example, the relevant United States field manual lists "surprises, ambushes, feigning attacks, retreats, or flights, simulating quiet and inactivity, use of small forces to simulate large units, transmitting false or misleading radio or telephone messages, deception of the enemy by bogus orders purporting to have been issued by the enemy commander (…) dummy guns and vehicles (…) dummy airfields", among many others [75].

Of notable importance for the boundary between ruses and perfidy are the examples given under Art. 37(1), namely "feigning of an intent to negotiate under a flag of truce or of a surrender", "feigning of an incapacitation by wounds or sickness", "feigning of civilian, non-combatant status", and "feigning of protected status by the use of signs, emblems or uniforms of the United Nations or of neutral or other States not Parties to the conflict". In addition, it is "prohibited to make use of the flags or military emblems, insignia or uniforms of adverse Parties while engaging in attacks or in order to shield, favour, protect or impede military operations" (Geneva Additional Protocol 1, Art. 39(2)). However, while using one's correct national insignia, it is permissible to remove unit identifications from uniforms [75].

Further relevant considerations concern the markings that must be placed on military equipment. For the case of unmanned aerial vehicles, Piatkowski [76] notes that "while the use of improper [physical] markings is prohibited and might in some circumstances be tantamount to perfidy, the use of [a] false IFF signature [electronic identification of friend vs. foe] is a lawful ruse of war" (though this is restricted to false military signatures, not civilian ones). More generally, authoritative legal commentary on air and missile warfare notes that "use of false military codes and false electronic, optical or acoustic means to deceive the enemy can be seen as a special case of lawful disinformation" [77]. Taking these considerations together, we postulate cases of lawful versus unlawful deception involving autonomous weapon systems, see Table 3.

With the exception of the case of a LAWS with no markings, all of the cases of unlawfulness in Table 3 relate to perfidy, that is, to killing, injuring, or capturing human enemy combatants. The first case, feigned surrender, could conceivably occur on a battlefield where both humans and machines are present and where one side announces a surrender. One would then expect the surrendering side to ensure that all LAWS are stood down. A surprise attack by a LAWS against human combatants under such circumstances

Table 3 Cases of lawful versus lawful deception involving autonomous weapon systems

Ruses of war (lawful deception)	Perfidy and other cases of unlawful deception
●Surprises and ambushes ●Simulating attack, retreat, or flight ●Simulating a larger or smaller force ●Simulating greater or lesser firepower ●Replacing a LAWS with a dummy ●Use of misleading electronic identification ●Use of bogus communications with other equipment or forces ●The LAWS has markings that identify its nationality, but unit markings are deliberately absent or incorrect ●An enemy LAWS was captured and repurposed, with its markings appropriately changed	●The LAWS appears to deliberately cease operation to lure humans and then attacks them ●The LAWS appears to be incapacitated to lure humans and then attacks them ●The LAWS is disguised as a civilian drone ●The LAWS uses the markings of the enemy, of a third country, or of an international organisation that is not a party to the conflict ●The LAWS has the markings of the Red Cross or an organisation of similar legal status ●The LAWS has no markings

would be perfidy. The second case is that incapacitation of a LAWS would be feigned for a treacherous purpose. One scenario could be that humans would be approaching a damaged enemy LAWS on the battlefield, e.g. for purposes of intelligence or scavenging of parts. It would be a perfidious act if the AWS reactivated and attacked these humans. However, perfidy does not apply to machines seeking to destroy, damage, or capture one another. Therefore, under the same scenario, if a robotic system were sent to analyse or scavenge parts from an apparently damaged LAWS, the LAWS could lawfully destroy that robotic system in a surprise attack.

Also, as machines are military equipment, there is no limit to lawful destruction of them. A battle exclusively between machines could lawfully be one of full attrition, namely until all machines of the losing side have been destroyed. The notion that machines would deliberately cease operation and allow capture by other machines is also not legally required, though conceptually feasible. In that case, the fate of the machines that allow capturing themselves could lawfully include their complete destruction. This implies that there would be no incentive for LAWS designers to allow for the LAWS to be captured by the LAWS of the enemy. Allowing for capture by human combatants of the opposing force would also provide no advantage, such that one should expect LAWS designers to ensure that LAWS engaged in a losing battle would at some point be able to decide to flee (assuming communications with their human commanders are lost), with the aim of mitigating military losses. Additionally, LAWS designers may wish to ensure that a LAWS that is unable to flee can destroy any sensitive on-board equipment and information, in case the enemy would seek to capture it for analysis, reverse engineering, or scavenging of parts. However, such self-destruct functions would have to avoid perfidious acts, e.g. deliberately detonating while being inspected by humans.

In the case of machines accepting capture by humans, rules on perfidy hold in favour of the humans. For the case of humans surrendering to a LAWS, the LAWS would have to respect rules on perfidy. This implies that an LAWS that is engaged in battle lawfully against human targets would necessarily be able to recognise an act of surrender and be capable of acting accordingly. If the LAWS is not capable of processing a surrender autonomously, it would have to at minimum, cease firing, report back to its human commanders, and await further instructions or human intervention. Conversely, humans who have surrendered to a LAWS would be allowed to betray the confidence of the LAWS at any later point in time in order to damage, destroy, or capture it. If control over the LAWS is taken over remotely by a human operator who belongs to the forces of the LAWS, then the surrender could perhaps be argued to be towards the human operator, but mediated by the LAWS. However, a feigned surrender in that case could not lead to killing, injuring, or capturing the human operator, and therefore feigning surrender with the intent to destroy the LAWS would be lawful. By the same logic, human combatants should be allowed to feign injury or sickness to an LAWS, even if that LAWS is remotely operated, as part of a ploy to destroy that LAWS.

In a hypothetical future battle between opposing formations of LAWS, the laws of warfare suggest that some of the actions that amount to perfidy if directed at humans would be lawful if directed at other machines—notably feigning surrender and feigning incapacitation. Also, deceptive electronic identification would be lawful—but not deceptive physical markings. This odd set-up results from the evolution of the laws of warfare intended for piloted military aircraft and provides an incentive for opposing formations to resort to spying with visual means. This could be, for instance, sending out a small party of LAWS that would need to get within visual range of the opposing force, close enough to recognise their markings. In addition, all of the ruses of war that are currently lawful would be lawful for LAWS as well. In sum, there is a wide scope for future LAWS to engage in deceptive practices, lawfully, on the battlefield. Therefore, states will undoubtedly seek to have such capabilities built into their future LAWS to achieve military advantage. How can LAWS capable of deception be reconciled with existing NATO principles of responsible use?

We will use the concepts of accuracy, secrecy, and trust as developed in the previous section to guide our reflections. We posit that for a LAWS to be able to practise lawful deception, it would have to fulfil, at a minimum, the following criteria:

1. The technical ability to carry out deception (from a computational perspective).
2. A high degree of accuracy in distinguishing:
   (a) Between friend, foe, and third party persons or objects (so as to engage and deceive only adversaries).
   (b) Between enemy objects that may be lawfully engaged, and those that may not (to avoid perfidy).
3. A high degree of secrecy, consistent with the US/NATO principle of reliability (which includes security).
4. An exceptionally high degree of trust between the LAWS and its human commanders, operators, and team mates.

The first criterion, the technical ability to carry out deception, could relate to "theory of mind" approaches. Humans who seek to deceive other humans use a theory of mind regarding their targets to induce them into making incorrect inferences. For example, the target may be induced into movements or actions that increase their vulnerability to attacks that the deceiving party is able to deploy. A theory of mind approach could also apply for machines trying to deceive other machines. If the attacking machine could extract or somehow derive the algorithms that control its target, it could anticipate the target's actions, and then stimulate the target into taking actions it prefers. From this perspective, the secrecy concept introduced earlier is therefore of extreme importance.

However, existing deep learning applications which are able to generate deceptive behaviour do not have theory of mind characteristics. It is rather through brute force iterative learning that the algorithm arrives, after millions of attempts, at a successful course of action which may include a deceptive move. The algorithm in such a case has no semantic understanding of its actions. Whether this poses an ethical problem may remain an open question. Provided the algorithm is constrained to avoid unlawful types of deception, one may make the determination that no unethical acts are carried out.

7 Conclusion

In our review of the NATO Principle of Explainability and Traceability, we explored the scope for acceptable limitations to both explainability and traceability. Three arguments were developed for further consideration, namely accuracy, secrecy, and trust. As noted, black box AI approaches, such as deep learning, may ensure only a limited degree of explainability and traceability. However, this need not come at the expense of sufficient accuracy, secrecy, and trust. Also, one should relate the notions of accuracy, secrecy, and trust back to the existing NATO principles of responsible use. Specifically, accuracy and secrecy should be viewed as fitting under the NATO principle of reliability. The principle of reliability was generally intended to refer to an AI system's technical ability to perform as intended, therefore including high accuracy. The reliability principle also makes an explicit mention of security. That mention was intended by its drafters to refer, notably, to an AI system being robust to various forms of electronic attacks that would cause it to malfunction or to reveal technically important information. Lastly, trust is addressed under the NATO principle of governability, which includes a commitment to "appropriate human–machine interaction". That wording was intended to include relevant work on ensuring trust between AI systems and their human operators or collaborators. Our suggestion for both practitioners and scholars would be to highlight the concepts of accuracy, secrecy, and trust within the existing NATO principles, and the use of these concepts to support the practical definition of what may constitute a sufficient fulfilment of the principle of explainability and traceability.

In our exploration of the possible implications of LAWS capable of deception, we have highlighted key types of operational behaviour that would be consistent or on the contrary inconsistent with IHL. A clear distinction can be made between LAWS actions against humans and other machines in a way that favours human agents, with the laws of warfare implying that certain actions, such as feigning incapacitation only to subsequently attack, would be considered perfidy directed at humans, but would be considered lawful ruses of war when directed at other machines. This clarification highlights that a significant range of ruses of war can be legally employed by LAWS, particularly against other machines. For an AWS to practise lawful deception, we propose four necessary criteria: the LAWS has the technical ability to carry out deception; the LAWS can accurately distinguish between adversary and non-adversary assets and persons; that the LAWS can act with secrecy in consistency with the US/NATO ethical principle on Reliability, and that there is sufficient trust between the LAWS, its operating chain of command, and team-mates.

It is not yet clear how to operationalise the proliferating sets of AI principles. Just as states continue to define, and approach, LAWS in different ways, it is likely that they will derive their own interpretations on how NATO principles should be best employed. This represents a challenge whereby states who take a strict, more restrictive approach to the ethical use of LAWS do so to their relative strategic detriment. Our discussion on deception in warfare highlights open questions about how LAWS may be employed to act in relation to humans, or other machines, in ways that represent legal deception in some cases, and perfidy in others. Extensive context-specific guidance will be required to operationalise principles thoroughly and sufficiently, in a way which minimises unethical activity. Correct implementation will take time, and significant learning-by-doing to determine how to apply principles in practice. These considerations should be included in the development of testing, evaluation, validation and verification (TEVV) of LAWS, ensuring such systems are employed in such a way that is consistent with agreed principles. The next few years will likely see an increase in national and international (including NATO and UN) initiatives that attempt to face these operationalisation challenges in terms of designing adequate TEVV procedures or certification assurance measures. Given the complexity of LAWS' technical operations, a flexible application of "methodological standards" will be more productive than traditional comprehensive testing which will fail to predict, or capture, all of the scenarios the system will face in a battlefield context.

Human judgement will be required throughout LAWS processes, from system design to governance, to ensure that LAWS are employed in a way that remains consistent with principles and IHL. The challenges of MHC in fully autonomous systems can be mitigated in part by explicit and documented designations to human responsibility throughout the decision-making cycle. The NATO Principles of Responsible Use provide a framework to inform human judgement in this way.

With evidence that the DoD and NATO are focusing their efforts on operationalising AI in warfare, there are several additional policy considerations for the employment LAWS. Militaries and defence organisations will need to invest significant resources, both financially and in terms of skilled expertise, to achieve adequate responsible use of LAWS (and AI systems in defence more broadly). Assessing the application of these principles to LAWS requires significant legal, policy and governance expertise as well as technical contributors. At a policy level, broader questions remain about how states will adopt AIs and deploy LAWS in conflict, and how far principles will be enforced in conflict scenarios. Beyond the questions of responsible use of AI, policymakers must face the trend of greater technological dependence both in terms of maintaining their security stances through rapid adoption, but in terms of encouraging responsible, and therefore more stable, use. In focusing on the latter objective, this paper highlights the importance of principles as a mechanism to address legal and ethical challenges associated with LAWS.

Funding The research of Ertan was supported by the EPSRC and the UK Government as part of the Centre for Doctoral Training in Cyber Security at Royal Holloway, University of London (EP/P009301/1). No other funds, grants, or other support was received.

Data availability In this article, we do not analyse or generate any datasets, because our work proceeds within a theoretical and ethical approach.

Declarations

Conflict of interest As a corresponding author, I confirm that Christie, Adomaitis, and Ertan held private company affiliations at the time of writing, but the research has been conducted with the full spirit of academic freedom with no commercial benefit to the affiliated entities. The views expressed by authors in this publication are their own and do not constitute the official position or policy of their affiliated organisations.

Ethical approval As a corresponding author, I confirm that no human subjects were involved in the study.

Informed consent As a corresponding author, I confirm that all authors consent to the article submission.

References

1. Taddeo, M., McNeish, D., Blanchard, A., Edgar, E.: Ethical principles for artificial intelligence in national defence. Philos Technol (2021). https://doi.org/10.1007/s13347-021-00482-3
2. Morgan, F.E., Boudreaux, B., Lohn, A.J., Ashby, M., Curriden, C., Klima, K., Grossman, D.: Military applications of artificial intelligence: ethical concerns in an uncertain world. Santa Monica, CA: RAND Corporation, 2020. https://www.rand.org/pubs/research_reports/RR3139-1.html
3. Scharre, P., Hawley, J., Schulman, L.D., McCarthy, M., Horowitz, M.C.: "Autonomous weapons and operational risk ethical autonomy project." (2016).
4. Taddeo, M., Blanchard, A.: A comparative analysis of the definitions of autonomous weapons. Sci. Eng. Ethics (2021). https://doi.org/10.1007/s11948-022-00392-3
5. Arkin, R.C.: Governing lethal behavior in autonomous robots. Chapman and Hall/CRC (2009)
6. Wagner, M.: The dehumanization of international humanitarian law: legal, ethical, and political implications of autonomous weapon systems. Vanderbilt J. Transnatl Law **47**, 1371 (2014)
7. US Department of Defense. (2020, 24 February): *DOD Adopts Ethical Principles for Artificial Intelligence, U.S. Department of*

Defense [Press Release]. https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/. Accessed 12 Feb 2022.
8. Defense Innovation Board. 2019: AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. Accessed 21 Feb 2022.
9. Defense Innovation Board.: AI principles: recommendations on the ethical use of artificial intelligence by the Department of Defense - supporting document. Defence Innovation Board (DIB). https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF (2019). Accessed 21 Feb 2022.
10. NATO.,: Summary of the NATO artificial intelligence strategy. Retrieved from https://www.nato.int/cps/en/natohq/official_texts_187617.html (2021, October 22). Accessed 6 Jan 2022.
11. Stanley-Lockman, Z., Christie, E.H.,: An Artificial Intelligence Strategy for NATO. *NATO Review.* Retrieved from https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html (2021, 25 October). Accessed 21 Feb 2022.
12 Wagner, M.: Taking humans out of the loop: implications for international humanitarian law. J. Law Inf. Sci. (2011). https://doi.org/10.5778/JLIS.2011.21.Wagner.1
13. Grut, C.: The challenge of autonomous lethal robotics to international humanitarian law. J. Confl. Secur. Law **18**(1), 5–23 (2013)
14. Arkin, R.: Lethal autonomous systems and the plight of the non-combatant. In: The political economy of robots, pp. 317–326. Palgrave Macmillan, Cham (2018)
15. Jackson, A.L., Kuenzli, K.D.: Something to believe in: aligning the principle of honor with the modern battlefield. Natl. Secur. Law J. **6**, 35 (2018)
16. Bathaee, Y.: The artificial intelligence black box and the failure of intent and causation. Harv. J. Law Technol. **31**(889), 891–892 (2018)
17. Lubell, N., Cohen, A.: Strategic proportionality: limitations on the use of force in modern armed conflicts. Int. Law Stud. **96**(1), 6 (2020)
18. Kilcup, J.: Proportionality in customary international law: an argument against aspirational laws of war. Chic. J. Int. Law **17**(1), 8 (2016)
19. Ali, S.: Coming to a battlefield near you: quantum computing, artificial intelligence, and machine learning's impact on proportionality. Santa Clara J. Int. Law **18**, 1 (2020)
20. Hua, S.S.: Machine learning weapons and international humanitarian law: rethinking meaningful human control. Georget. J. Int. Law **51**, 117 (2019)
21 Müller, V.C.: Ethics of artificial intelligence. In: Elliott, A. (ed.) The Routledge social science handbook of AI. Routledge, London (2021)
22. Arkin, R. C.: Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture. In: Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pp. 121–128. (2008, March)
23. Margulies, P.: Autonomous weapons in the cyber domain: balancing proportionality and the need for speed (April 22, 2020). Roger Williams Univ. Legal Studies Paper No. 201, Available at SSRN: https://ssrn.com/abstract=3582580
24. Schuller, A.L.: At the crossroads of control: the intersection of artificial intelligence in autonomous weapon systems with international humanitarian law. Harv. Natl. Secur. J. **8**, 379 (2017)
25. Coffin, A.M.: Lethal autonomous weapons systems: can targeting occur without ethical decision-making? United States Naval War College (2019)
26. Boulanin, V., Davison, N., Goussac, N., Carlsson, M.P.: Limits on autonomy in weapon systems: identifying practical elements of human control. SIPRI (2020)
27. Boulanin, V., Bruun, L., Goussac, N.: Autonomous weapon systems and international humanitarian law: identifying limits and the required type and degree of human-machine interaction. ICRC (2021)
28. Horowitz, M.C.: The ethics and morality of robotic warfare: assessing the debate over autonomous weapons. Daedalus **145**(4), 25–36 (2016)
29. Dremliuga, R.: General legal limits of the application of the lethal autonomous weapons systems within the purview of international humanitarian law. J. Politics Law. **13**, 115 (2020). https://doi.org/10.5539/jpl.v13n2p115
30. Taddeo, M., Blanchard, A.: A comparative analysis of the definitions of autonomous weapons. Sci. Eng. Ethics **28**(5), 37 (2021)
31. Boardman, M., Butcher, F.: An exploration of maintaining human control in AI enabled systems and the challenges of achieving it. NATO: Technical report (2019)
32. Amoroso, D., Tamburrini, G.: Autonomous weapons systems and meaningful human control: ethical and legal issues. Curr. Robot Rep. **1**, 187–194 (2020). https://doi.org/10.1007/s43154-020-00024-3
33. Scholz, J., Galliott, J.: The humanitarian imperative for minimally-just AI in weapons, p. 57. Lethal Autonomous Weapons: Re-Examining the Law and Ethics of Robotic Warfare (2020)
34. Sassoli, M.: Autonomous weapons and international humanitarian law: advantages, open technical questions and legal issues to be clarified. Int. Law Studies/Naval War Coll. **90**, 308–340 (2014)
35. McDougall, C.: Autonomous weapon systems and accountability: putting the cart before the horse. Melb. J. Int. Law **20**(1), 58–87 (2019)
36. Sullins, J.P.: Deception and virtue in robotic and cyber warfare. In: Taddeo, M., Floridi, L. (eds.) The ethics of information warfare, pp. 187–201. Springer, Cham (2014)
37 Chelioudakis, E.: Deceptive AI machines on the battlefield: do they challenge the rules of the law of armed conflict on military deception? SSRN Electron. J. (2017). https://doi.org/10.2139/ssrn.3158711
38 Anderson, K., Reisner, D., Waxman, M.C.: Adapting the law of armed conflict to autonomous weapon systems. International Law Studies, US Naval War College (2014)
39. Sharkey, A.: Autonomous weapons systems, killer robots and human dignity. Ethics Inf. Technol. **21**(2), 75–87 (2019). https://doi.org/10.1007/s10676-018-9494-0
40. US Department of Defense.: Establishment of the Chief Digital and Artificial Intelligence Officer. [Memorandum]. https://media.defense.gov/2021/Dec/08/2002906075/-1/-1/1/MEMORANDUM-ON-ESTABLISHMENT-OF-THE-CHIEF-DIGITAL-AND-ARTIFICIAL-INTELLIGENCE-OFFICER.PDF (2021a, 8 December). Accessed 10 Jan 2022
41. US GGE Statement.: "Reviewing potential military applications of emerging technologies in the areas of lethal autonomous weapons systems". U.S. Statement at the GGE on laws during the discussion of agenda item 5(D) (2021, 5 August)
42. US Department of Defense.: Autonomy in weapons systems (DoD Directive 3000.09). https://www.hsdl.org/?abstract&did=726163 (2012, November 12): Accessed 10 Jan 2022.
43. US GGE Statement.: "Human-machine interaction in the development, deployment and use of emerging technologies in the area of lethal autonomous weapons systems". CCW/GGE.2/2018/WP.4. https://undocs.org/CCW/GGE.2/2018/WP.4 (2018, 28 August). Accessed 8 Feb 2022.
44. UK Ministry of Defence.: Defence artificial intelligence strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artif

icial_Intelligence_Strategy.pdf (2022, 16 March). Accessed 4 Nov 2022.

45. GCHQ.: Pioneering a new national security - the ethics of artificial intelligence. https://www.gchq.gov.uk/artificial-intelligence/index.html#footnotes (2021). Accessed 2 Nov 2022
46. UK Ministry of Defence.: Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in defence. Available at https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence (June 15, 2022). Accessed 4 Nov 2022.
47. UK MoD.: UK weapons review. development, concepts and doctrine centre. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/507319/20160308-UK_weapon_reviews.pdf (2016, August). Accessed 10 Feb 2022.
48. French Defense Ministry.: "Defence ethics committee - opinion on the integration of autonomy into lethal weapon systems." https://cd-geneve.delegfrance.org/IMG/pdf/defence_ethics_committee_-_opinion_on_the_integration_of_autonomy_into_lethal_weapon_systems.pdf?2423/17d8f6beb2f5c9caa9c9168c53c24a91d9d32513 (2021, 9 April). Accessed 2 Nov 2022.
49. Jeangène-Vilmer, J-B.: A French Opinion on the Ethics of Autonomous Weapons. *War on the Rocks.* Available at https://warontherocks.com/2021/06/the-french-defense-ethics-committees-opinion-on-autonomous-weapons (2021, 2 June). Accessed 8 Feb 2022.
50. Ministry for Europe and Foreign Affairs.: (Online). 11 principles on lethal autonomous weapons systems (LAWS). Retrieved from https://www.diplomatie.gouv.fr/en/french-foreign-policy/united-nations/multilateralism-a-principle-of-action-for-france/alliance-for-multilateralism/article/11-principles-on-lethal-autonomous-weapons-systems-laws. Accessed 21 Feb 2022.
51. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
52. China.: China's comments on the working recommendations of the group of governmental experts on laws. https://documents.unoda.org/wp-content/uploads/2021/06/China.pdf (2021). Accessed 2 Nov 2022.
53. Kania, E.: China's strategic ambiguity and shifting approach to lethal autonomous weapons systems. https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems (2018). Accessed 2 Nov 2022.
54. HRW.: Killer robots: negotiate new law to protect humanity. https://www.hrw.org/news/2021/12/01/killer-robots-negotiate-new-law-protect-humanity (2021). Accessed 2 Nov 2022.
55. Jankowski, Dominik P.: Russia and the technological race in an era of great power competition. Center for International and Strategic Studies. https://www.csis.org/analysis/russia-and-technological-race-era-great-power-competition *(2021).* Accessed 2 Nov 2022.
56. Russian Federation.: "Considerations for the report of the group of governmental experts of the high contracting parties to the convention on certain conventional weapons on emerging technologies in the area of lethal autonomous weapons systems on the outcomes of the work undertaken in 2017–2021". Available at: https://documents.unoda.org/wp-content/uploads/2021/06/Russian-Federation_ENG1.pdf (2021). Accessed 28 Oct 2022.
57. Kokkinos, Matthew A.: Global governance of autonomous weapon systems: the Russia case study. MALD capstone requirement. The Fletcher School of Law and Diplomacy – Tufts University. https://sites.tufts.edu/fletcherrussia/files/2020/05/The-Global-Governance-of-AWS-Russia-Case-Study.pdf (2020). Accessed 21 Feb 2022.
58. GGE Statement.: Report of the 2018 session of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems. CCW/GGE.1/2018/3. https://undocs.org/en/CCW/GGE.1/2018/3 (2018, 23 October). Accessed 8 Feb 2022.
59. GGE Statement.: Final report. CCW/MSP/2019/9. https://undocs.org/CCW/MSP/2019/9 (2019, 13 December). Accessed 8 Feb 2022.
60. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach Intell. **1**, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
61. Wehmeier, S., Raaz, O.: Transparency matters: the concept of organizational transparency in the academic discourse. Public Relat. Inq. **1**(3), 337–366 (2012). https://doi.org/10.1177/2046147X12448580
62. Searle, J.: Minds, brains, and programs. Behav. Brain Sci. **3**(3), 417–424 (1980). https://doi.org/10.1017/S0140525X00005756
63. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., Yu, H.: Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705. http://www.jstor.org/stable/26600576 (2017)
64. Wang, D., Yang, Q., Abdul, A., Lim, B. Y.: Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1–15) (2019)
65. Lundén, J., Koivunen, V.: Deep learning for HRRP-based target recognition in multistatic radar systems. In: 2016 IEEE Radar Conference (RadarConf) (pp. 1–6). IEEE. (2016)
66. Holland Michel, A.: The black box, unlocked: predictability and understandability in military AI. Ginebra, United Nations Institute for Disarmament Research, disponible en https://unidir.org/publication/black-box-unlocked (2020)
67. Figdor, C.: Pieces of mind: the proper domain of psychological predicates. Oxford University Press (2018)
68. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: a review of machine learning interpretability methods. Entropy **23**(1), 18 (2021). https://doi.org/10.3390/e23010018
69. Devine, S. M., Bastian, N. D.: Intelligent systems design for malware classification under adversarial conditions. arXiv preprint arXiv:1907.03149, https://doi.org/10.48550/arXiv.1907.03149 (2019)
70. Mahmood, K., Gurevin, D., van Dijk, M., Nguyen, P.H.: Beware the black-box: on the robustness of recent defenses to adversarial examples. Entropy **23**(10), 1359 (2021). https://doi.org/10.3390/e23101359
71 Winikoff, M.: Towards trusting autonomous systems. In: International workshop on engineering multi-agent systems, pp. 3–20. Springer, Cham (2017)
72. Lord, K.M.: The perils and promise of global transparency: why the information revolution may not lead to security, democracy, or peace. Suny Press (2007)
73. Lewis, M., Yarats, D., Dauphin, Y.N., Parikh, D., Batra, D.: Deal or no deal? End-to-end learning for negotiation dialogues. arXiv preprint arXiv:1706.05125, https://doi.org/10.48550/arXiv.1706.05125 (2017)
74 Caddell, J.W.: Deception 101-primer on deception. Army War College Strategic Studies Institute, Carlisle Barracks, PA (2004)
75. U.S. Department of the Army: The law of land warfare, pp. 27–10. Department of the Army Field Manual, FM (1956)
76. Piątkowski, M.: The markings of military aircraft under the law of aerial warfare. Mil Law Law War Rev **58**(1), 63–84 (2020)
77. Harvard School of Public Health. Program on Humanitarian Policy, Conflict Research, Program on Humanitarian Policy, and Conflict Research at Harvard University: HPCR manual on international law applicable to air and missile warfare. Cambridge University Press (2013)

78. Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States.: Principles and good practices on emerging technologies in the area of lethal autonomous weapons systems. Joint submission to the United Nations Group of Governmental Experts on Lethal Autonomous Weapon Systems". 7 March (2022)
79. Argentina, Costa Rica, Guatemala, Kazakhstan, Nigeria, Panama, Philippines, Sierra Leone, State of Palestine, Uruguay.: "Proposal: roadmap towards new protocol on autonomous weapons systems". 7 March (2022)
80. Christie, Edward Hunter and Amy Ertan.: NATO and artificial intelligence. In: Romaniuk, S.N., Manjikian. M. (Eds.) Routledge Companion to Artificial Intelligence and National Security Policy. Routledge, Forthcoming (2022)
81. NATO.: "Summary of NATO's autonomy implementation plan". Available at https://www.nato.int/cps/sn/natohq/official_texts_208376.htm (2022). Accessed 4 Nov 2022
82. NATO.: "NATO's data and artificial intelligence review board". Available at https://www.nato.int/cps/fr/natohq/official_texts_208374.htm (2022). Accessed 4 Nov 2022
83. Allen, Gregory C, "DOD Is Updating Its Decade-Old Autonomous Weapons Policy, but Confusion Remains Widespread".: Center for strategic and international studies. Available at https://www.csis.org/analysis/dod-updating-its-decade-old-autonomous-weapons-policy-confusion-remains-widespread (2022). Accessed 4 Nov 2022
84. Congressional Research Service.: Defense primer: U.S. policy on lethal autonomous weapon systems. Available at https://crsreports.congress.gov/product/pdf/IF/IF11150 (2022). Accessed 4 Nov 2022.
85. Embassy of the People's Republic of China in the United States of America.: "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)". http://us.china-embassy.gov.cn/eng/zgyw/202201/t20220113_10492264.htm (2022). Accessed 4 Nov 2022.
86. Nadibaidze, A.: Great power identity in Russia's position on autonomous weapons systems. Contemp Sec Policy **43**(3), 407–435 (2022). https://doi.org/10.1080/13523260.2022.2075665

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