European Integration Studies, Volume 20, Number 2 (2024), pp. 189-211.
https://doi.org/10.46941/2024.2.8

KAJA KOWALCZEWSKA*

Human oversight and risk-based approach to artificial intelligence: What does the Artificial Intelligence Act have in common with discussions about lethal autonomous weapon systems?**

ABSTRACT: The regulation of artificial intelligence (AI) is a pressing global concern, with various regulatory bodies aiming to foster innovation while safeguarding humanity's interests. This article synthesises perspectives on AI regulation in civilian and military domains, highlighting common ethical foundations and legal proposals. Emphasising the European Union's ethical community as delineated by fundamental rights, it explores the Artificial Intelligence Act and debates on lethal autonomous weapon systems within the Convention on Certain Conventional Weapons. By analysing the overlap between civilian and military ethics, the article argues for a shared objective: promoting innovation while upholding human dignity through robust regulations that ensure human oversight and a risk-based approach. The article contends that the consensus on substantive issues regarding military AI regulation is imminent, but its formalisation through legal means may lag behind.

KEYWORDS: Artificial Intelligence, Ethics, European Union, Lethal Autonomous Weapon Systems, Human Oversight, Risk-Based Approach.

1. Introduction

Various regulatory bodies worldwide are formulating regulations pertaining to the utilisation of artificial intelligence (AI), each drawing upon its specialised expertise. Amidst this diversity, a shared objective emerges: the regulation of AI to foster innovation for the betterment of humanity. This leads to the following question: what exactly does this objective mean, and how can it be guaranteed through legal regulations? This article summarises

* This article came about within the framework of Academic Excellence Hub – Digital Justice Center carried out under Initiative of Excellence – Research University at the University of Wrocław, Poland. kaja.kowalczewska@uwr.edu.pl.
** The research and preparation of this study was supported by the Central European Academy.

the main perspectives regarding the regulation of AI in civilian and military applications, focusing on identifying common ground, particularly concerning ethical foundations and the associated legal proposals.

One of the primary challenges associated with ethics is the fact that ethical concerns, which underpin and anticipate legal norms, vary significantly across states. However, within the European Union (EU), which is not only an economic entity but also a community founded on shared values, it can be argued that such an ethical community is delineated by values that are legally safeguarded and enshrined as fundamental rights.[1] Accordingly, this article examines the general-purpose AI regulation outlined in the Artificial Intelligence Act (AIA)[2] as well as the ongoing discourse surrounding lethal autonomous weapon systems (LAWS) within the Convention on Certain Conventional Weapons (CCW) forum. The discussion focuses on the positions advanced by EU member states in these discussions as exemplified in the two-tier approach and draft articles. The selection of these two overarching AI categories is predicated on the premise that legal norms governing peacetime and armed conflict, despite their apparent dichotomy and to a limited extent, share common ethical principles. I contend that the prohibition of certain use cases incompatible with the requirement of public conscience (unacceptable risk) and the insistence on human responsibility that cannot be delegated to AI-based machines (human oversight) represent such paramount considerations. In civilian AI applications, the requisites of public conscience are grounded in values such as democracy, the rule of law, and human rights.[3] Conversely, in military contexts, they derive from the paradigm of international humanitarian law (IHL), which entails balancing the principles of humanity and military necessity. Thus, the overarching objective in both realms of AI applications is the promotion of technological innovation for the collective benefit of humankind, guided by ethical considerations rooted in principles aimed at protecting human dignity.

With the recent adoption of the AIA and the emergence of other global initiatives,[4] the conversation surrounding AI regulation has shifted

[1] Wouters, 2020, pp. 11–38.
[2] European Parliament, 2024.
[3] Załucki and Miraut, 2021.
[4] Responsible Artificial Intelligence in the Military (REAIM), 2023; United Nations Secretary-General, 2023; United Nations Educational, Scientific and Cultural Organization, n.d.; Organisation for Economic Co-operation and Development, n.d.

from being taboo to becoming one of the most prominent subjects of discussion.[5] While discussions within the CCW forum have persisted for over a decade, tangible outcomes remain elusive. The AIA, particularly on a regional level, has clearly delineated several unacceptable risks associated with potent AI models. This juncture may signify a crucial moment, especially concerning matters of warfare, with states still deliberating the acceptability of various autonomous weapons and the conditions under which they may be employed. This article proposes that consensus on substantive matters is imminent but that formalisation through legal regulation may remain distant.

Regulations, as exemplified by those articulated in the AIA, adopt a risk-based methodology to define the parameters of acceptable AI applications, demarcating the thresholds beyond which certain uses are considered unacceptable. Traditionally, regulations governing security and warfare have been distinct from those governing civilian affairs and peace. However, AI is unique in its capacity to gradually blur these boundaries, as in the need to address the bias issue.[6] This phenomenon is particularly apparent in technologies with dual-use capabilities, serving both military and law enforcement purposes. Art. 2(3) of the AIA underscores this convergence, thereby blurring the distinction between security-related and civilian-focused regulatory concerns. Consequently, debates surrounding LAWS bear similarities to those concerning the deployment of social scoring or real-time biometric classification systems. In both cases, the central issue is how to delineate the ethical and moral boundaries of technological integration within societal and armed conflict contexts. Thus, the risk-based approach is increasingly permeating discussions on military applications of AI.

First, this article examines the AIA, emphasising its ethical foundation, risk-based methodology, and human oversight, including the exclusion of military applications. Second, the article explores debates on LAWS, highlighting concerns about unacceptable risks and the necessity of human oversight as emerging common points, followed by a conclusion.

[5] Ramos et al., 2024, p. 34.
[6] Bode, 2024.

2. Artificial Intelligence Act: The landmark law on general purpose artificial intelligence

The European Parliament adopted the AIA on 13 March 2024, marking a significant milestone in technology governance.[7] Negotiated with the member states in December 2023, the regulation garnered widespread recognition as a landmark law, signalling a unified stance on advancing a new governance model rooted in technology. While the official version of the EU regulation remains pending, indications suggest that substantial amendments are unlikely, with the anticipated revisions being primarily cosmetic. Following three years of negotiations, the EU has emerged as a trailblazer in the legal regulation of civilian AI applications. The following section will offer an exposition of the AIA's ethical principles and a synthesis of the adopted risk-based approach, concluding with a discussion of why, in principle, the AIA does not extend to military AI applications.

The AIA represents a comprehensive legislative endeavour aimed at fostering technological innovation while safeguarding fundamental rights, particularly in contexts where highly impactful AI models pose risks. This alignment with fundamental rights protection is not unexpected, given the EU's adherence to the Charter of Fundamental Rights[8] and the commitment of its member states to the European Convention for the Protection of Human Rights and Fundamental Freedoms.[9] These states have pledged to uphold high standards of human rights protection amidst evolving technological and economic landscapes.

In the AIA, numerous references underscore the importance of values, such as health protection, safety, fundamental rights, democracy, the rule of law, and environmental sustainability. Moreover, the text of the AIA reflects a cohesive approach towards these emerging values, which have surfaced in discussions regarding broader AI applications.[10] This is why the fundamental prerequisite outlined in the AIA is the establishment of trustworthiness in AI.[11] Central to this notion is the concept of "human-

[7] Members of the European Parliament. overwhelmingly approved the Act, with 523 votes in favour, 46 against, and 49 abstentions.
[8] Charter of Fundamental Rights of the European Union, 2000.
[9] Convention for the Protection of Human Rights and Fundamental Freedoms, 1950.
[10] Trustworthy AI, human agency and oversight, and traceability and explainability.
[11] Díaz-Rodríguez et al., 2023.

centricity", wherein AI is envisioned as a tool serving the interests of people, with the overarching objective of enhancing human well-being.[12] In order to fulfil this mission, AIA draws from the meta-framework of ethical consideration that preceded the regulatory effort and should be presented as a normative prerequisite of the legal regulation.

2.1 Ethical underpinnings

The AIA builds upon the foundational work of the AI High-Level Expert Group, which established seven non-binding ethical principles for AI aimed at ensuring trustworthiness and ethical integrity in AI development and deployment.[13] The preamble of the AIA declares that efforts should be made to integrate these principles into the design and utilisation of AI models wherever feasible. Furthermore, they are posited as fundamental components for the creation of codes of conduct consistent with AI regulations. The recommendation extends to all stakeholders, encompassing industry players, academic institutions, civil society organisations, and standards bodies, who are encouraged to adopt these ethical principles as they craft voluntary best practices and standards. Thus, these principles constitute essential pillars that should underpin any forthcoming regulatory framework governing AI within the EU.

Paramount consideration is accorded to the principles governing human agency and oversight. Within this framework, AI systems must be conceptualised and operationalised as instruments subservient to human interests while upholding fundamental tenets of human dignity and personal autonomy. Such systems must be engineered to operate within parameters amenable to human control and supervision, thereby ensuring alignment with ethical imperatives.[14]

Furthermore, the imperatives of technical robustness and safety require AI systems to be resilient against operational exigencies and impervious to external manipulations aimed at subverting their intended utility. This necessitates the development and deployment of AI technologies with robust mechanisms capable of withstanding adversities and thwarting illicit attempts to exploit or alter their functionalities for unlawful ends.

---

[12] Kowalczewska, 2021a, pp. 465–486.
[13] Stix, 2021, p. 15.
[14] Puscas, 2022.

Adherence to established regulatory frameworks governing data protection and privacy rights is of paramount importance in the realm of privacy and data governance. Therefore, AI systems must adhere rigorously to stipulated norms, ensuring data processing of impeccable quality and integrity, thereby safeguarding privacy rights and preserving data sanctity.[15]

Transparency, as a guiding principle, requires the elucidation of AI systems' inner workings, affording stakeholders insights into the dynamics of human–AI interactions. This entails furnishing users with comprehensive information regarding the operational modalities, capabilities, and limitations of AI systems, thereby fostering informed decision-making and engendering a culture of accountability.[16]

Additionally, the principles of diversity, non-discrimination, and fairness mandate the equitable treatment of individuals irrespective of their demographic attributes. Artificial intelligence systems are enjoined to promote inclusivity, gender equality, and cultural diversity while eschewing discriminatory practices or biases that contravene established legal standards.

Moreover, the ethical imperative of societal and environmental well-being necessitates the sustainable development and deployment of AI technologies. It is imperative that AI innovations not only serve to ameliorate human welfare but also mitigate adverse environmental impacts, thereby ensuring the perpetuation of societal equilibrium and ecological harmony.[17]

Finally, the principle of accountability mandates that AI systems be subject to stringent mechanisms of oversight and redress. This entails delineating clear lines of responsibility and establishing robust frameworks for recourse in the event of malfeasance or adverse outcomes attributable to AI operations.[18]

This ethical meta-framework largely aligns with other soft-law instruments developed in various AI-oriented forums. As demonstrated later, the framework was applied extensively in the AIA but also found significant resonance in discussions concerning LAWS.

[15] Michel, 2021.
[16] Michel, 2020.
[17] United Nations Institute for Disarmament Research, 2015.
[18] Anand and Deng, 2023.

2.2 Risk-based approach

An integral aspect of the AIA is the delineation of AI-based systems. As articulated in Art. 3(1) of AIA, these systems are characterised as:

> …machine-based systems designed to operate with varying levels of autonomy, capable of exhibiting adaptability upon deployment, and with the capacity to infer from inputs received how to generate outputs such as predictions, content, recommendations, or decisions that may impact the physical or virtual environment.

Given the expansive scope of applications encompassed by the AIA, this definition involves a broad scope and has, consequently, been subject to criticism.[19] However, it underscores a crucial attribute of AI systems (similar to the definition-related discussion regarding LAWS)—namely, their capacity for inference-making, along with varying degrees of autonomy from human intervention and the potential to execute actions without direct human involvement. Naturally, such autonomous action entails inherent risks, which are addressed in the provisions of the AIA.

In developing the AIA, a risk-based approach was adopted,[20] wherein AI systems were categorised into four distinct levels based on the risks they pose to fundamental rights: unacceptable, high, limited, and minimal (or no) risk. The AIA assigns specific obligations to providers and users based on the level of risk associated with the AI system. Of particular significance for this article are the first two categories, which delineate prohibited uses of AI and those necessitating human oversight.

2.2.1 Unacceptable risks and prohibitions of certain artificial intelligence systems

Art. 5 of the AIA prohibits placing AI systems on the market, putting them into service, or using them in several specific scenarios. These scenarios include the employment of manipulative or deceptive techniques, exploitation of vulnerabilities, implementation of social scoring systems for natural persons, deployment of biometric categorisation systems, and real-

[19] Ruschemeier, 2023, pp. 361–376.
[20] Key Issue 3: Risk-Based Approach - EU AI Act, n.d.

time remote biometric identification of individuals in publicly accessible spaces for law enforcement purposes.

The AI systems categorised under this prohibition are considered harmful to individuals and are, therefore, completely barred from use within the EU space, with only limited exceptions for specific law enforcement purposes. The prohibited applications primarily involve scenarios in which continuous surveillance could lead to discrimination, substantial violations of privacy and freedom of movement, or other significant harms. Although why these specific systems are deemed contrary to democratic values is not elaborated, this decision is based on certain principles and falls within the realm of the political discretion vested in lawmakers. Similarly, the international community anticipates analogous decisions within discussions on LAWS, wherein states should interpret the fundamental principles of IHL and decide on the extent of the LAWS regulation.

2.2.2 High-risk and human oversight

The high-risk category of AI systems, as defined in Art. 6 of the AIA and further detailed in Annex III, requires AI systems to meet two conditions to qualify for classification within this group. First, the AI system must be subjected to the EU harmonisation legislation outlined in Annex I. Second, the system must undergo a third-party conformity assessment according to the same legislation.

The broad definition of high-risk AI applications encompasses a spectrum of schemes perceived as risky owing to their potential to cause significant harm across multiple domains, including health, safety, fundamental rights, the environment, democracy, and the rule of law. Examples of high-risk AI applications can be found in various sectors, such as critical infrastructure; education; employment; essential private and public services like healthcare and banking; certain law enforcement systems; migration and border management; and justice and democratic issues like election integrity. These examples highlight the diverse contexts in which high-risk AI implementations may pose substantial threats, thus warranting heightened scrutiny and regulation under the AIA.

Under this regulation, such systems are permitted on the market but are subject to a comprehensive set of conditions aimed at the provision of trustworthy AI. These include the implementation of a robust risk-management system (Art. 9), adherence to stringent data management and

governance practices (Art. 10), the maintenance of thorough technical documentation (Art. 11), and the establishment of comprehensive record-keeping protocols (Art. 12). Furthermore, transparency and informed instructions for use must be provided (Art. 13), and effective human oversight must be ensured throughout the system's lifecycle (Art. 14). Additionally, AI systems should maintain an appropriate level of accuracy, robustness, and cybersecurity (Art. 15). The affected individuals are entitled to obtain clear and meaningful explanations from the deployer regarding the AI system's role in decision-making processes and the key elements of the decisions made (Art. 86). This last provision also reflects a commitment to transparency and explainability for AI-based processes.

Within the high-risk category of AI systems, significant emphasis is placed on human oversight.[21] The AIA mandates human oversight through three key pillars: the provision of appropriate human–machine interface tools; the objective of preventing or minimising risks; and the introduction of oversight measures tailored to the risks, autonomy level, and use context of the high-risk AI system. These measures can be integrated by the provider or implemented by the deployer. Through these pillars, the individual responsible for executing human oversight is expected to possess the capacity to understand the relevant capabilities of the AI system and effectively monitor its operation to detect and address any anomalies. They should maintain the awareness of automation bias and interpret outputs generated by the AI system appropriately. Additionally, they must be able to exercise the authority required to withdraw or override decisions made by the AI system and halt the operation of the AI system by pressing a stop button under safe conditions.

However, criticism has been raised regarding the AIA's approach to human oversight, suggesting that it focuses on procedural guidelines for AI system providers and lacks substantive guidance on the effectiveness of this oversight.[22] Additionally, concerns have been voiced about the considerable freedom granted to AI system providers, particularly regarding the circumstances triggering oversight.[23] It is argued that a decision of such significance, embodying the essence of human oversight, should, at the very least, be accompanied by a set of guidelines formulated by lawmakers dedicated to safeguarding fundamental rights.

---

[21] Key Issue 4: Human Oversight - EU AI Act, n.d.
[22] Laux, 2023.
[23] Enqvist, 2023, p. 534–535.

Nevertheless, this human oversight framework serves as a solid starting point that can be enhanced by targeted regulation, best practices, and technical designs developed within the respective fields of AI system deployment. When examining the debate surrounding LAWS, a similar challenge arises in regulating human oversight in a qualitative manner without being overly restrictive or narrow. Moreover, the approaches differ slightly. In the military setting, there is a greater emphasis on user oversight, particularly by military commanders rather than providers. While awaiting more detailed guidance, it is important to acknowledge that the regulation of human oversight in the AIA represents a commendable yet preliminary step in establishing a legal framework for trustworthy AI.

2.3 Exclusion of military purposes

Any secondary law adopted in the EU, such as a regulation like the AIA, must be based on primary law. Primary law is where member states determine the allocation of competences among EU institutions and retain certain areas as sovereign competences. National security matters, including defence, are among those areas that member states have chosen to retain as their sole responsibility under Art. 4(2) and Chapter 2 of Title V of the Treaty on European Union

Given the primary objectives of the EU's existence, matters pertaining to world peace and security have traditionally fallen within the realm of public international law rather than EU law. This is underscored in the preamble to the AIA, which acknowledges that 'public international law is therefore the more appropriate legal framework for the regulation of AI systems in the context of the use of lethal force and other AI systems in the context of military and defence activities'. Consequently, the EU consistently excludes applications related to national security and warfare from the scope of its laws.[24] The provisions of the AIA are consistent with this approach.

According to Art. 2(3) of the AIA, the regulation explicitly excludes national security matters from its scope, irrespective of whether these tasks are carried out by public or private entities. Notably, it specifies that the AIA does not apply to AI systems when they are marketed, used, or exploited solely for military, defence, or national security purposes or when their outputs are utilised exclusively for such purposes within the EU, even if the systems themselves do not operate within its territory.

[24] Compare Recital 16 of the General Data Protection Regulation.

This somewhat cryptic formulation can be elucidated by considering the interpretation provided in the context of Recital 24 of the preamble. It clarifies that if the primary purpose of placing or using an AI system is for a military, defence, or national security application, then it falls outside the scope of the AIA. However, if such a system is subsequently used outside its military purpose temporarily or permanently, such as for civilian, humanitarian, or law enforcement purposes, it falls back within the scope of the AIA. The same rule applies to AI systems designed for mixed purposes (both military and civilian), wherein only the civilian-purpose use falls under the scope of the AIA.

Under this convoluted regulation, the AIA is not applicable when an AI system is intended for military purposes or is used by any entity for military purposes. It appears that the drafters of the AIA considered the dual-use nature of AI systems but also framed the exceptions in the use-case language (rather than technology-type language) that is used consistently throughout the AIA. They adopted this approach to exclude military actors engaged in military operations while including civilian uses of AI systems originally conceived for military purposes.

Nevertheless, I contend that the exclusion of AI systems developed for military purposes is not based primarily on the distinct ethical underpinnings of such military-oriented AI systems but is based on the formal issue of the competence division between EU institutions and member states. This is reflected in the normative referral of this issue from the realm of EU law to public international law. Indeed, discussions about military AI systems are ongoing in forums like the CCW, in which individual member states and the EU, with its competence as an observer, are actively participating. Furthermore, they present positions that are in line not only with IHL but also with the ethical principles expressed in the AIA.

3. Discussions about lethal autonomous weapon systems in the Convention on Certain Conventional Weapons forum

Discussions within the CCW, initiated by coalitions of non-governmental organisations such as Stop Killer Robots, have continued for over a decade. Despite the adoption of various formats, including informal expert meetings and gatherings of government experts, these deliberations have yet to progress towards a negotiation of a legally binding international instrument. While civil society holds expectations of such progress, particularly

concerning LAWS operating without meaningful human control and potentially creating an accountability gap,[25] the likelihood of achieving this goal has been minimal from the outset. Furthermore, the prevailing international security landscape further diminishes the possibility of such a solution in the foreseeable future.[26] However, these discussions have seen some progress, and I contend that officially embracing the positions outlined below would be perceived by Stop Killer Robots and the states supporting this position as a triumph and, fundamentally, a recognition of their demands.

While there is no universally accepted single definition of LAWS, states generally concur in principle that LAWS encompass weapon systems that, once activated, can identify, select, and engage targets with lethal force without further intervention by an operator.[27] By translating this definition into the language of the AIA, it can be inferred that LAWS are AI systems intended for military purposes that exhibit adaptability post-deployment. These systems are capable of inferring, from received inputs, how to generate outputs such as decisions regarding the identification and selection of military targets, which may influence the physical environment through engagement with military targets (including people or objects), potentially resulting in serious incidents. Although current discussions are considered to pertain to lethal AI applications, a detailed analysis reveals that some positions are broader, encompassing decision-support systems and other autonomous or remotely piloted means of warfare that do not pose risks similar to those posed by LAWS.[28] To narrow the scope of this discussion and focus on the most critical applications (i.e. those with lethal consequences), the discussion will concentrate on issues related to LAWS.

There is a widespread consensus that all developed and employed means of warfare must adhere to IHL.[29] This means that as a state's right to develop and deploy weapons is limited, the weapons must be utilised in compliance with the fundamental principles of IHL, including distinction, proportionality, precautions, and the prohibition against causing unnecessary suffering or superfluous injury.[30] However, I argue that this

[25] Human Rights Watch, 2012; Human Rights Watch, 2015.
[26] Puscas, 2023.
[27] CCW, 2023a.
[28] Bo and Dorsey, 2024.
[29] CCW, 2019.
[30] Kowalczewska, 2021b, pp. 88–103.

assertion may not be adequate to comprehensively regulate LAWS. I contend that the intrinsic nature of AI-based decision-making on matters of life and death, without clear human accountability, warrants examination by lawmakers to determine its acceptability, particularly in light of established customs, principles of humanity, and the mandates of public conscience (Martens clause).[31] This requires states to declare their stance on the acceptable level of risk to fundamental rights, especially the right to life, within the context of armed conflict. Consequently, they should adopt a risk-based approach, akin to that outlined in the AIA, by explicitly prohibiting certain uses of LAWS and regulating high-risk LAWS more tightly with a set of mitigating measures. This perspective is increasingly evident in statements presented at the CCW. In the sections below, I will focus on two propositions recently put forth by several EU member states to highlight convergent points and demonstrate the gradual emergence of this approach in discussions.

3.1. The two-tier approach

The so-called "two-tier approach" was proposed in July 2022 by a group of European states, comprising Finland, France, Germany, the Netherlands, Norway, Spain, and Sweden.[32] The states proposed a possible structure of recommendations for measures related to a normative and operational framework.

3.1.1. Unacceptable risks

The core concept underlying this approach posits that 'autonomous weapons systems that cannot comply with IHL are effectively prohibited and should neither be developed nor used, necessitating further efforts to implement this commitment at the national level'. This seemingly straightforward and legally obvious assertion is elaborated upon in a more nuanced manner, providing insight into which types of AI systems, according to a two-tier approach, are deemed unacceptable and warrant regulation. The former category comprises LAWS that operate entirely beyond human control or a responsible chain of command; the latter pertains to all other types of LAWS. From the delineation of these LAWS categories, one can infer that, according to these states, LAWS lacking both a responsible chain of

[31] Ibid., p. 228.
[32] CCW, 2022.

command and appropriate human control inherently contravene IHL and should be de jure prohibited.

This statement can be contrasted with the AIA's classification of unacceptable risks, but there is a significant difference: LAWS, as AI systems under scrutiny, are normatively embedded within the IHL framework, which offers some direction on their acceptability. By contrast, civilian applications are governed according to human rights standards. Weapons law and IHL embody legal frameworks that are more robust than the civilian regulation of AI as they focus on specific military actions such as targeting. Thus, regulations are stricter for states deploying such systems in combat than in the broader and less-explored commercial settings. States can discern which systems pose unacceptable risks within established normative frameworks. In civilian AI, per the AIA, these systems generate risks incongruent with rights to privacy, human dignity, and protection from discriminatory practices. In the military sphere, attention is drawn to risks that would lead to an accountability gap. This disparity in approach reflects the distinct ethical foundations of both frameworks, which prioritise different values during peace and war.

3.1.2. Human oversight

Another aspect of the two-tier approach is its emphasis on human oversight, akin to the AIA. For LAWS other than those classified as unacceptable, this oversight entails appropriate human control and a responsible chain of command.[33] The author states define appropriate human control as encompassing human oversight over the entire lifecycle of LAWS, including the development, deployment, and utilisation phases. This oversight should ensure that LAWS operate predictably, enabling humans to ascertain their compliance with legal, political, and operational standards and ensuring the explainability of their operations. During the development stage, human control should involve the testing, certification, and legal review of LAWS to evaluate their reliability and predictability. During the deployment phase, human control should manifest in the establishment of rules of engagement and a delineation of the mission objectives, target types, and spatial and temporal constraints while monitoring the system's reliability and usability within this context. Finally, during utilisation, humans should retain decision-making authority over the use of force,

[33] For a critical approach to this framework, see Article 36, 2023.

which encompasses a scope broader than a mere attack. It includes the ability to approve any significant changes in mission objectives, maintain communication links, and deactivate the system, although the technical feasibility of the latter action is deemed optional by the author states.

The second condition entails maintaining human responsibility and state accountability throughout the lifecycle. This aligns with the ethical imperative wherein a human should always be held responsible for the actions of machines, thus heeding the call to address the accountability gap. This condition is considered satisfied through the implementation of several measures, including the development of LAWS-specific doctrines and procedures and the provision of adequate training on LAWS for human decision-makers and operators. It also entails ensuring that the responsible chain of human command encompasses human accountability for the creation and validation of rules of operation, use, and engagement as well as decision-making regarding deployment. This approach implies the introduction of after-action review measures. It also advocates for maintaining the accountability framework, which involves reporting, investigation, prosecution, and disciplinary procedures in cases of grave breaches of IHL due to the use of LAWS.

3.2. Draft articles

A more robust approach, known as the "Draft Articles on Autonomous Weapons", was introduced in May 2023 by a coalition of states, including EU member state Poland, along with Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States.[34] This approach outlines autonomous systems in Art. 1 that should not be developed owing to their conflict with IHL principles. The subsequent articles focus on detailed regulatory measures to ensure the effective implementation of fundamental IHL principles: Art. 3 emphasises distinction, Art. 4 addresses proportionality, and Art. 5 highlights precautions. The final article, Art. 6, pertains to the accountability regime. To maintain consistency in the analysis, the presentation of the draft articles will follow the previous logic of a risk-based approach and human oversight indicators.

[34] CCW, 2023b.

3.2.1. Unacceptable risks

The proposing states assert that certain AI systems, by virtue of their design, pose unacceptable risks and are, therefore, incompatible with IHL. These systems include those that cause harm to civilians and civilian objects by targeting them, spreading terror, or consistently leading to disproportionate collateral damage. The articulation within the draft articles unequivocally establishes that only LAWS deliberately designed to contravene IHL principles are deemed unlawful. This assertion, while legally evident and akin to the two-tier approach, also reflects a pragmatic understanding of the nature of weapons and the regulatory framework governing armed conflict. It acknowledges the inherent purpose of weapons to cause harm while emphasising that only attacks on civilians and civilian objects that are intentional and disproportionate can be classified as war crimes. Consequently, states involved in deliberations regarding the acceptability of LAWS operate under the premise that AI-based weapon systems possess the capability to make critical life-and-death decisions. However, they converge with Stop Killer Robots on the second condition, concerning human responsibility.

The draft articles explicitly specify that LAWS operating outside the responsibility framework of commanders or their operators are considered unacceptable under this proposal. This stance aligns with the two-tier approach by prohibiting LAWS that would operate without human responsibility attached to their actions or those designed in contravention of IHL principles, as outlined in Annex I.

Therefore, it appears that states supporting the two-tier approach and draft articles generally align with the main argument of Stop Killer Robots. However, they differ in their willingness to be legally bound by this standard.

3.2.2. Human oversight

States supporting the draft articles contend that all other LAWS categories should be designed to foresee and manage their effects during attacks according to the principles of distinction and proportionality. In pursuit of this objective, they delineate various sets of risk-mitigating measures aimed at upholding fundamental IHL principles and establishing an effective accountability framework. During development, these measures should

include testing, evaluation, and legal review, along with limit-setting regarding target types, duration, geographical scope, and scale (e.g. self-destruct, self-deactivation, or self-neutralisation mechanisms), as well as addressing automation and unintended bias. Furthermore, the draft articles, offering a more detailed framework than the two-tier approach, underscore the significance of certain principles. These include the reliance on LAWS in good faith, taking into account the information available at the time of the use of force and exercising due diligence in adhering to IHL principles, as elucidated in Articles 3–5.

The draft articles establish an accountability framework within the broader context of implementing IHL and additional LAWS-specific measures. The former encompasses measures such as education and training on IHL, a responsible chain of human command and control, the development of domestic legislation, international reporting mechanisms, and appropriate investigations, which may entail accountability for personnel. The latter involves easily understandable human–machine interfaces and controls, guidance and training for personnel on the appropriate use of LAWS, and specific rules of engagement and other military documentation relevant to military operations.

Hence, the draft articles emphasise that LAWS should conform to the overarching IHL framework, encompassing all conventional rules. Moreover, they delineate specific measures targeted at ensuring the effective implementation of these norms, particularly in light of the unique characteristics of AI systems.

4. Conclusion

In this article, I aimed to demonstrate the emerging normative consensus on the need for human oversight and risk-based approaches for AI regulation. As examples, I used the AIA, covering a broad group of general-purpose AI systems, and discussions on military applications of AI in the form of LAWS. Although the examples involved different regimes of factual situations (i.e. peacetime and wartime), I attempted to show that a limited ethical anchorage could be commonly found across EU member states (as well as other states).

Utilising a risk-based methodology facilitates the identification of AI systems whose operations contravene core legal norms, such as those governing democracy, human rights, or IHL, thereby warranting their

prohibition. Conversely, for AI systems categorised under lower risk levels, tailored regulatory measures can be instituted to mitigate societal exposure to their potentially adverse ramifications. Within these deliberations, ethical principles such as human agency; technical robustness; and reliability, predictability, transparency, explainability, and human accountability have assumed central importance, resonating across discussions concerning the AIA and LAWS. These ethical precepts constitute integral components of a broader normative framework that remains indispensable in the AI discourse. The imperative now is to meticulously situate these principles within the specific operational context and milieu of the pertinent use case.

Furthermore, the discussion highlighted the imperative to ensure human oversight, particularly in instances where risks are deemed acceptable but are elevated. This underscores a reluctance to entrust decision-making to AI systems in contexts of ethical significance, such as in critical services, judicial proceedings, and the employment of force. The operationalisation of such oversight ought to be predicated upon a cohesive comprehension of procedural imperatives (what actions to undertake and when) and qualitative mandates (the rationale behind actions), which should be delineated not solely by ethical precepts but also be enshrined within legal regulatory frameworks.

Finally, the most notable disparity between the two cases concerns regulation. While the AIA serves as a directive targeting economic entities, mandating compliance for profit generation within the EU, its adoption is relatively straightforward compared with the negotiation and implementation of a multilateral arms treaty. Nonetheless, I posit that, if EU member states are committed to upholding the normative values that are fundamental to the EU, they should actively articulate, in a legally binding manner, the unacceptable risks posed by AI in armed conflict, thereby affirming their adherence to fundamental ethical principles such as human dignity. However, the current geopolitical landscape, underscored by Russia's aggression against Ukraine in 2022, has engendered reluctance among states to embrace new arms control commitments, and some have even contemplated withdrawing from existing commitments. Consequently, while the calls from Stop Killer Robots for a ban on such weapons may be unavailing at present, it is hoped that the positions articulated in the two-tier approach and draft articles will suffice to prevent the development, deployment, or utilisation of the most hazardous AI systems—those endowed with full unsupervised autonomy and lethal capabilities.

Bibliography

[1] Anand, A., Deng H. (2023) *Towards Responsible AI in Defence: A Mapping and Comparative Analysis of AI Principles Adopted by States*. [Online]. Available at https://unidir.org/publication/towards-responsible-ai-in-defence-a-mapping-and-comparative-analysis-of-ai-principles-adopted-by-states/ (Accessed: 1 July 2024).

[2] Bo, M., Dorsey, J. (2024) *Symposium on Military AI and the Law of Armed Conflict: The 'Need' for Speed – The Cost of Unregulated AI-Decision Support Systems to Civilians.* Opinio Juris. [Online]. Available at: https://opiniojuris.org/2024/04/04/symposium-on-military-ai-and-the-law-of-armed-conflict-the-need-for-speed-the-cost-of-unregulated-ai-decision-support-systems-to-civilians/ (Accessed: 10 April 2024).

[3] Bode, I. (2024) The problem of algorithmic bias and military applications of AI. *Humanitarian Law & Policy Blog*. [Online]. Available at: https://blogs.icrc.org/law-and-policy/2024/03/14/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai/ (Accessed: 29 March 2024).

[4] Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, H., Herrera, F. (2023) 'Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation' *Information Fusion* 99, p. 101896; https://doi.org/10.1016/j.inffus.2023.101896.

[5] Enqvist, L. (2023) 'Human oversight' in the EU artificial intelligence act: what, when and by whom?' *Law, Innovation and Technology* 15(2), pp. 508–535; https://doi.org/10.1080/17579961.2023.2245683.

[6] Kowalczewska, K. (2021a) *Unia Europejska wobec autonomicznych systemów śmiercionośnych broni (LAWS) – znacząca ludzka kontrola jako fundament wiarygodnej sztucznej inteligencji* in Fischer, B., Pązik, A., Świerczyński, M. (eds.) *Prawo sztucznej inteligencji i nowych technologii*, Wolters Kluwer Polska, pp. 465–486.

[7] Kowalczewska, K. (2021b) *Sztuczna inteligencja na wojnie. Perspektywa MPHKZ. Przypadek autonomicznych systemów śmiercionośnej broni.* Wydawnictwo Naukowe Scholar.

[8] Laux, J. (2023) 'Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act', *AI & SOCIETY*; https://doi.org/10.2139/ssrn.4377481.

[9] Michel, A. (2020) *The Black Box, Unlocked: Predictability and Understandability in Military AI.* [Online]. Available at: https://unidir.org/publication/the-black-box-unlocked/ (Accessed: 1 July 2024).

[10] Michel, A. (2021) *Known Unknowns: Data Issues and Military Autonomous Systems*. United Nations Institute for Disarmament Research. [Online]. Available at: https://unidir.org/publication/known-unknowns (Accessed: 1 July 2024).

[11] Puscas, I. (2022) *Human-Machine Interfaces in Autonomous Weapon Systems*. [Online]. Available at: https://unidir.org/publication/human-machine-interfaces-in-autonomous-weapon-systems/ (Accessed: 1 July 2024).

[12] Puscas, I. (2023) *AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures.* [Online]. Available at: https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/ (Accessed: 1 July 2024).

[13] Ramos, G., Squicciarini, M., Lamm, E. (2024) 'Making AI Ethical by Design: The UNESCO Perspective', *Computer*, 57(2), pp. 33–43; https://doi.org/10.1109/MC.2023.3325949.

[14] Ruschemeier, H. (2023) 'AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal', *ERA Forum*, 23(3), pp. 361–376; https://doi.org/10.1007/s12027-022-00725-6.

[15] Stix, Ch. (2021) 'Actionable Principles for Artificial Intelligence Policy: Three Pathways', *Science and Engineering Ethics* 27(1), p. 15; https://doi.org/10.1007/s11948-020-00277-3.

[16] Wouters, J. (2020) *From an economic community to a union of values: The emergence of the EU's commitment to human rights*. in Wouters, J. et al. *The European Union and Human Rights*. Oxford University Press. pp. 11-38. https://doi.org/10.1093/oso/9780198814191.003.0002.

[17] Załucki, M., Miraut, M. (2021). *Artificial intelligence and human rights.* Dykinson.

[18] Article 36. (2023) *Completely outside human control?* [Online]. Available at: https://article36.org/wp-content/uploads/2023/03/Completely-outside-human-control.pdf (Accessed: 29 March 2024).

[19] CCW (2022) *Working paper submitted by Finland, France, Germany, the Netherlands, Norway, Spain, and Sweden to the 2022 Chair of the Group of Governmental Experts (GGE) on emerging technologies in the area of lethal autonomous weapons systems (LAWS)*. [Online]. Available at: https://documents.unoda.org/wp-content/uploads/2022/07/WP-LAWS_DE-ES-FI-FR-NL-NO-SE.pdf (Accessed: 1 July 2024).

[20] CCW (2023a) *Draft articles on autonomous weapon systems – prohibitions and other regulatory measures on the basis of international humanitarian law ("IHL")*. [Online]. Available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.4_US_Rev2.pdf (Accessed: 1 July 2024).

[21] CCW (2023b) *Non-exhaustive compilation of definitions and characterizations*. [Online]. Available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_CRP.1_0.pdf (Accessed: 1 July 2024).

[22] Convention for the Protection of Human Rights and Fundamental Freedoms (1950).

[23] Charter of Fundamental Rights of the European Union (2000).

[24] European Parliament (2024) *Artificial Intelligence Act. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. [Online]. Available at: https://www.europarl.europa.eu/RegData/seance_pleniere/textes_adoptes/definitif/2024/03-13/0138/P9_TA(2024)0138_EN.pdf (Accessed: 29 March 2024).

[25] Human Rights Watch. (2012) *Losing humanity: the case against killer robots*. Amsterdam Berlin: Human Rights Watch. [Online]. Available at: https://www.hrw.org/sites/default/files/reports/arms1112_ForUpload.pdf (Accessed: 29 March 2024).

[26] Human Rights Watch. (2015) *Mind the Gap: The Lack of Accountability for Killer Robots*. [Online]. Available at: https://www.hrw.org/sites/default/files/reports/arms0415_ForUpload_0.pdf (Accessed: 1 July 2024).

[27] OECD (n.d.) *AI-Principles Overview*. [Online]. Available at: https://oecd.ai/en/ai-principles (Accessed: 1 July 2024).

[28] REAIM (2023) *REAIM 2023 Call to Action.* [Online]. Available at: https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action (Accessed: 1 July 2024).

[29] UNESCO (n.d.) *Ethics of Artificial Intelligence*. [Online]. Available at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics (Accessed: 28 August 2023).

[30] UNIDIR (2015) *The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values*. [Online]. Available at: https://unidir.org/publication/the-weaponization-of-increasingly-autonomous-technologies-considering-ethics-and-social-values/ (Accessed: 1 July 2024).

[31] UNSG. (2023) *Secretary-General's remarks to the Security Council on Artificial Intelligence*. [Online]. Available at: https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence. (Accessed 1 July 2024).

[32] (n.d.) *Key Issue 3: Risk-Based Approach - EU AI Act.* [Online]. Available at: https://www.euaiact.com/key-issue/3 (Accessed: 29 March 2024).

[33] (n.d.) *Key Issue 4: Human Oversight - EU AI Act*. [Online]. Available at: https://www.euaiact.com/key-issue/4 (Accessed: 29 March 2024).