Digital Society (2025) 4:23
https://doi.org/10.1007/s44206-025-00181-y

ORIGINAL PAPER

A Risk-Based Regulatory Approach To Autonomous Weapon Systems

Alexander Blanchard[1] · Claudio Novelli[2] · Luciano Floridi[2,3] · Mariarosaria Taddeo[4]

Received: 2 April 2024 / Accepted: 12 March 2025

Abstract

International regulation of autonomous weapon systems (AWS) is increasingly conceived as an exercise in risk management. This requires a shared approach for assessing the risks of AWS. This paper presents a structured approach to risk assessment and regulation for AWS, adapting a qualitative framework inspired by the Intergovernmental Panel on Climate Change (IPCC). It examines the interactions among key risk factors—determinants, drivers, and types—to evaluate the risk magnitude of AWS and establish risk tolerance thresholds through a risk matrix informed by background knowledge of event likelihood and severity. Further, it proposes a methodology to assess community risk appetite, emphasising that such assessments and resulting tolerance levels should be determined through deliberation in a multistakeholder forum. The paper highlights the complexities of applying risk-based regulations to AWS internationally, particularly the challenge of defining a global community for risk assessment and regulatory legitimisation.

Keywords Autonomous weapons systems · Artificial intelligence · Risk threshold · Predictability problem · Governance · Regulation

Alexander Blanchard and Claudio Novelli contributed equally to this work and shared as first authors.

✉ Mariarosaria Taddeo
mariarosaria.taddeo@oii.ox.ac.uk

1 Stockholm International Peace Research Institute (SIPRI), Signalistgatan 9, 169 72, Solna, Sweden

2 (DEC), Yale University, 85 Trumbull St., New Haven, CT 06511, USA

3 Department of Legal Studies, University of Bologna, IT Digital Ethics Center, Via Zamboni, 27/29, Bologna, Italy

4 Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

Published online: 03 April 2025

1 Introduction

International regulation of autonomous weapon systems (AWS)[1] for use in armed conflict is increasingly conceived as an exercise in risk management, namely, what assessment and mitigation measures are needed for reducing the risks of non-compliance with international humanitarian law (IHL) (Bhuta and Pantazopoulos 2016; Geiss, 2016; Roff and Moyes, 2016, p. 4; McFarland, 2022, p. 403). For example, this is the approach taken by the International Committee of the Red Cross (ICRC) (2021).[2] However, for international debate to advance, a framework is required to assess the risks of AWS. Currently, no such framework exists. While risk-based regulatory approaches for AI are emerging in other domains, such as with the European Union regulation on AI (the AI Act), these do not extend to the use of military AI.[3] In addition, some literature identifies AWS risk types (Scharre, 2016; Laird, 2020; Johnson, 2020), but currently, there is no framework for describing the interaction(s) among these types nor an approach for a comparative assessment of risks against ethical and legal standards.

A risk-based approach would be key to understanding risks and steering decision-making to address the challenges associated with using AWS in armed conflict. This approach would facilitate a balanced and reasoned regulation of technologies that, while presenting inherent risks, could potentially yield benefits (Etzioni, 2018), and would also be pertinent for managing emerging types of risks.[4]

In the case of AWS, risks related to the unpredictability of their behaviour during deployment are particularly relevant (Taddeo, 2024; Taddeo et al., 2022).[5] In this case, risk assessment would help address uncertainty problems and enhance the ability to evaluate and quantify the unpredictability associated with AWS behaviour

[1] We define an autonomous weapon system as "…an artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal, or set of goals, within its dynamic operating environment and without the direct intervention of another agent and may also be endowed with some abilities for changing its own transition rules without the intervention of another agent, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object or a human being) and to this end is able to identify, select and attack the target without the intervention of another agent is an AWS. Once deployed, AWS can be operated with or without some forms of human control (in, on or out the loop)" (Taddeo and Blanchard, 2022, p. 15).

[2] Similarly, guiding principle (g) affirmed by the UN CCW GGE (2019) states that: "Risk assessments and mitigation measures should be part of the design, development, testing and deployment of emerging technologies in any weapons systems." See also the May 2023 report by Automated Decision Research (2023), 'Convergences in state positions on human control', pp. 10–12.

[3] There exist candidate frameworks for managing the risks of AI systems in the civil domain (Baybutt, 2014; NIST, 2023), but the distinctive character of armed conflict — such as differing prerogatives and risk thresholds for combatants and noncombatants — necessitate a specialized model for risk evaluation and management.

[4] Emerging risk can be defined as: "[…] the likelihood of loss, i.e. the probability of a certain consequence to occur in specific time and space under specified or insufficiently specified conditions" (Flage & Aven, 2015, p. 62).

[5] The unpredictability of an AI system is not boundless but limited by the system affordances – the set of hardware and software specifications that determine the range of possible actions of a machine.

during deployment.[6] The critical aspect of the predictability problem with AWS is its impact on assessing potential violations of relevant international law, particularly IHL. This difficulty extends to protecting civilians and combatants, with repercussions on normative approaches to AWS. An example of these approaches is that of the ICRC, which recommends that:

> Unpredictable autonomous weapon systems should be expressly ruled out, notably because of their indiscriminate effects […] This would best be achieved with a prohibition on autonomous weapon systems that are designed or used in a manner such that their effects cannot be *sufficiently* understood, predicted and explained (ICRC, 2021, p. 2 - italics added).

However, implementing such a prohibition would face challenges as delineating between predictable and unpredictable AWS is not straightforward since (un)predictability of AWS behaviour exists as a spectrum ranging from entirely predictable to completely unpredictable.[7]

At the same time, the ICRC's appeal for a *sufficient* level of understanding highlights the importance of establishing a comprehensive framework for evaluating the predictability of AWS. This framework could help facilitate the identification of systems with a requisite level of predictability for meeting ethical and legal standards, thereby separating them from those that do not, with the possibility of feeding into an emerging two-tier regulatory approach (ICRC, 2022; van den Boogaard, 2024; Bruun, 2024).

Moreover, predictability should not be considered in isolation. The acceptability of AWS, including its regulatory treatment, should be determined through a comprehensive assessment that includes predictability, among other risk factors. For example, an AWS — or a component thereof — that is highly unpredictable but has a negligible actual impact might be considered acceptable under certain conditions.

Following these considerations, this paper proposes a risk-based regulatory approach to AWS. Such an approach offers a basis for a common understanding of AWS risk and a structured method for quantifying various uncertainties associated with AWS, including but not limited to (un)predictability. It can feed into emerging regulatory approaches and enable a transparent and scientifically sound examination of necessary trade-offs in the regulatory process.

The paper is structured as follows. Section 2 adapts a qualitative framework from the Intergovernmental Panel on Climate Change (IPCC) and applies it to AWS. This section shows how understanding the interactions among three key risk fac-

[6] The predictability problem refers to the limited certainty with which one can predict the behaviour of an AWS, and more broadly of AI systems, once deployed (Taddeo et al., 2022). Unpredictable systems are not a new issue. They are common in mathematics and physics, and limits on the ability to predict the outcomes of artificial systems have been proven formally since the 1950s (Moore, 1990; Musiolik & Cheok, 2021; Rice, 1956). Wiener and Samuel debated the predictability of AI systems in a famous exchange in 1960 (Wiener, 1960; Samuel, 1960).

[7] In real-world scenarios, neither of these two extremes is likely to be encountered: on one end are mechanistic systems with limited adaptability, and on the other, systems no military would (or ought) deploy due both to the excessive risk involved and lack of military effectiveness.

tors—determinants, drivers, and types—helps assess the overall risk magnitude of AWS (subsection 2.1). It also discusses setting a risk tolerance threshold using a risk matrix, enhanced by considering background knowledge of event likelihood and severity (subsection 2.2). Section 3 introduces a methodology for evaluating a community's risk appetite and its integration with established risk thresholds, noting that definitive risk appetite and tolerance levels should result from discussions in a legitimate multistakeholder forum. The paper acknowledges the challenges of applying risk-based regulation for AWS, highlighting the difficulty of defining a global community for risk assessment and legitimisation purposes, a task more daunting internationally than in domestic contexts (Peel, 2010). Section 4 concludes by summarising the main insights.

2 A Risk Framework for AWS: Risk Magnitude, Risk Appetite, and Background Knowledge

Risk-based policies can streamline governance interventions by setting priorities and objectives based on explicit criteria, such as (1) *risk magnitude* – the combination of harm likelihood and severity of consequences – and (2) *risk appetite* – the amount of risk that an organization or individual is willing to take in pursuit of their objectives.[8] These factors inform resource and cost distribution about risk management and allow for coping with uncertainties by making probabilistic predictions about potential hazards. Another criterion is (3) *background knowledge*, which refers to the robustness of the information underpinning a particular belief about the probability of a risk event occurring and its potential outcomes (Aven, 2015, 2017). Essential criteria for this assessment include the availability and relevance of data, the validity of assumptions, comprehension of the studied phenomena and processes, consensus among experts, and an evaluation of the foundational knowledge of the study.

The efficacy of risk-based governance strategies in establishing nuanced tolerance thresholds for AWS is significantly enhanced by the specificity of the three criteria outlined. To assess the risks associated with AWS, one must evaluate their potential positive and negative impact, and then balance this against the risk appetite of the concerned community, considering its values, goals, cultural practices, and expectations. Specific tolerance thresholds result from such a blend of external (magnitude) and internal (appetite) perspectives on risk. Risk tolerance thresholds aim in this context to inspire normative guidelines for regulating AWS, determining either their prohibition, or their broad or conditional acceptability, thus informing risk management measures.[9]

[8] A definition of risk tolerance understood as "risk appetite" of this kind can be found in the ISO Risk Management – Principles and Guidelines (2009).

[9] Similar frameworks can be found elsewhere. One example is the ALARP ('As Low As Reasonably Practicable') principle which serves as a normative standard within UK legislation for risk management, especially within industries critical to safety. This principle also plays a pivotal role in the UK health system's framework for risk management. It establishes three categorizations of risk tolerance: unacceptable risk, tolerable risk, and broadly acceptable risk (Baybutt, 2014; Abrahamsen et al., 2018).

To assess the risk magnitude of AWS, we use a risk framework originating from IPCC climate change risk policy reports and their subsequently expanded literature (Simpson et al., 2021). Climate change and AWS risks pose similar challenges, as both result in increasingly complex risk magnitudes due to the interplay of various factors and a significant reliance on specific contexts and stakeholders (Bhuta and Pantazopoulos 2016). Their associated risks represent global issues that require a unified international response.

2.1 The Climate Change Risk Framework for Qualifying Risk Magnitude of AWS

The IPCC framework identifies critical risk factors and their interactions for risk assessment of an event. According to the revised framework (Simpson et al., 2021), risk magnitude results from three sets of interactions between: (1) determinants, (2) drivers, and (3) types.[10] The first set refers to interactions between four risk determinants: (H) Hazard, (E) Exposure, (V) Vulnerability, and (R) Response. The second set refers to risk drivers, that is, the individual components of determinants. Risk drivers impact the overall risk magnitude by interacting in various ways, such as aggregating, compounding, or cascading.[11] The third set of interactions refers to how the risk magnitude of a specific event interacts with other risk types, each having its determinants and drivers. Risk types can be grouped into extrinsic risks and ancillary risks. We shall illustrate (non-exhaustively) the content of all these risk factors by applying them to the AWS scenario:

- **(*H*)** Hazard denotes any potential source or condition that could adversely affect individuals, things, values, or the environment. Several hazard drivers may have an impact on the risk magnitude of AWS. For example, the level of autonomy in determining how much human supervision is needed in operating the AWS poses potential risks due to malfunctioning or poorly specified reward functions. Multiple levels and shades of autonomy are pertinent to this discussion, such as whether autonomy pertains to system critical functions, like target identification and selection, or in manoeuvring and navigation (Boulanin et al., 2020; Longpre et al., 2022). Malfunctions can trigger a chain reaction, leading to multiple unintended engagements (Leys, 2018). Adaptive capabilities, e.g., through machine learning methods, may enable AWS to learn from past environmental interactions or present battlefield conditions, enhance performance, and adjust to changing conditions, but also increase overall unpredictability (Hua, 2019; Holland Michel, 2020; Schwarz, 2021; Taddeo and Blanchard, 2022). Target recognition capability, e.g., whether the use of AWS minimizes collateral damage and enhances adherence to IHL or, by contrast, undermines respect for IHL by increasing rates of noncombatant harm. Several hazard drivers associated with

[10] These three sets of interactions occur at levels of increasing complexity.

[11] These interactions cut across determinants, drivers, and risk types. Aggregate is when independent drivers collectively impact the risk magnitude; compounding is when a combination of drivers unidirectionally or bi-directionally affects the risk; cascading is when one driver sets off others, leading to a chain reaction of further drivers.

common AI vulnerabilities have significant consequences for AWS, such as model overfitting, adversarial attacks (Akhtar and Mian 2018), transfer learning problems (Weiss, Khoshgoftaar, and Wang 2016), and data poisoning (Biggio and Roli 2018). Payload is another critical hazard driver since the type and quantity of payload influences its danger level: e.g., an AWS equipped with a cluster munition is more dangerous than one equipped with munitions with a smaller blast radius. Other key hazard drivers of AWS include their operational range and duration – long-range AWS, carrying out missions hundreds or thousands of miles away from their control stations, present risks of prolonged engagement without direct human supervision due to reduced response time and diminished operator situational awareness (Boulanin et al., 2020; Horowitz, 2021).

- **(*E*)** Exposure denotes the inventory of items in reach of a hazard source. Exposure drivers for AWS can include the civilian population proximate to the operational area and combatants directly involved in the conflict (Nurick, 1945; McMahan, 2010). However, it can also encompass other military personnel, such as ground-based troops near AWS operations but not directly involved in their control or deployment, and personnel involved in supporting roles, such as logistics and maintenance (Blanchard and Taddeo, 2023). The potential harm from AWS can also extend to physical entities such as infrastructure – e.g., bridges and power-plants – or natural environments, including natural resources like groundwater, and digital ecosystems such as data systems. AWS can impact value-based assets, including those rooted in normative, e.g., Just War Theory) (Blanchard and Taddeo, 2022) and legal frameworks (e.g., IHL) (Anderson & Waxman, 2017; Sassoli, 2014) directly or through their material implications. For example, AWS deployment may violate the principle of distinction (Melzer, 2008), as algorithms for target identification might not reliably distinguish between combatants and noncombatants. Indirectly, the deployment of AWS without proper assessment could breach precautionary obligations, such as the principle of precautions in attack, detailed in Article 57 of the Additional Protocol I to the 1949 Geneva Conventions, requiring all conflict parties to take feasible measures to minimize harm to civilians and civilian objects (Winter, 2022; Thurnher, 2018).
- **(*V*)** Vulnerability denotes the attributes or circumstances that make the exposed elements susceptible to adverse effects when exposed to the hazard source. Vulnerability drivers relate to the exposed elements, i.e., individuals, things/environment, and values. For individuals, the vulnerability drivers may differ between combatants and noncombatants. For combatants operating AWS, vulnerabilities primarily arise from inadequate measures for handling AWS and preventing their malfunctioning, e.g., inadequate protection, insufficient training, unfamiliarity with AWS, and over-reliance. Noncombatants may face increased risks due to proximity to conflict areas, lack of information about and countermeasures to AWS, socioeconomic or coercive limitations preventing them from relocating, and difficulties in providing rescue or protection to specific groups. Social and existential factors act as vulnerability drivers, resulting in unequal impacts from the deployment of AWS. This approach underscores the necessity of acknowledging the increased risks borne by individuals with disabilities, marginalized communities, populations in the Global South, and women and girls, and incorporating

these considerations into assessments of AWS deployment impacts (Muñoz and Díaz 2021; Conway, 2020; Figueroa et al., 2023; Ramsay-Jones, 2019; Chandler, 2021; UN Women, 2023; Blanchard and Bruun, 2024). The dependence on assistive technologies or medical devices, susceptible to disruption by AWS collateral damage, can also heighten vulnerability (Quinn, 2021). Past exposure to armed conflict or weapon systems deployment can also heighten vulnerability amongst noncombatant populations. This can have potentially complex and far-reaching effects. For example, studies on remote warfare in places like Afghanistan and Pakistan have documented the psychological impact of drone presence on civilians (Edney-Browne, 2019). These effects are notably severe among marginalized and vulnerable groups, including children, who may cease attending school, and women, who have experienced increased rates of miscarriage due to the stress associated with the constant threat (Molyneux, 2021; Figueroa et al., 2023).

Vulnerability drivers regarding things and the environment may be the fragility of some (digital) infrastructures, e.g. their vulnerability to cyber-attacks, their strategic importance for populations, and the long-term susceptibility of some environments to damage. Lastly, ethical, and legal values could be at greater risk when there are weak political and/or judicial institutions, inefficient legal safeguards about AWS, and inadequately defined rules.

- **(*R*)** Response pertains to (pre-existing) strategies and measures that mitigate risk magnitude, revealing a community's resilience to specific risks. AWS response drivers encompass technological solutions and legal or governance protections (Molyneux, 2021). Technological drivers include fail-safe mechanisms (e.g., through function allocation) (Canellas & Haga, 2015), effective warnings, geofencing, remote deactivation, and operational duration limits. Response drivers include mechanisms for halting or suspending attacks to prevent breaches of distinction or proportionality principles and adjusting attack timing to minimize civilian exposure, such as programming AWS to prevent targeting in populated areas (Thurnher, 2018).

From the legal point of view, response drivers include incentives for adopting technical safeguards as well as governance measures, such as mandatory external controls and audits for AWS deployment (e.g., through oversight authorities), effective liability frameworks, certification protocols, transparency mandates, obligatory reporting, practices that foster trust in military operations (Roff and Danks 2018; Blanchard et al., 2024; Taddeo et al., 2022), standards for data quality, and regulatory sandboxes. Whether an AWS manufacturer or implementer complies with conformity assessments, for example, by obtaining certification from a recognized body, it typically indicates a lower risk profile, either broadly or in specific instances (Novelli et al., 2024). Nonetheless, due to the lack of comprehensive AWS regulations, only a few measures effectively mitigate other risk factors. Despite this, AWS ongoing development and use are already influencing perceptions of what is technically and ethically feasible, outstripping deeper societal and political debate on what should be normatively acceptable (Bode and Huelss 2018).

Against this background, such technical or legal response drivers in any context necessitate revaluating the overall risk associated with deploying AWS.

The overall risk level associated with AWS should also be considered in relation to its interaction with extrinsic and ancillary risk types. Several extrinsic factors could amplify the overall risk magnitude. A critical, extrinsic risk involves regulatory and liability uncertainties, specifically the lack of clear and cohesive regulations governing AWS design, development, deployment, and usage. The absence of unified rules and coordination can lead to an increased risk of civilian harm or violations of IHL. This issue becomes particularly acute in multinational operations involving multiple states, each adhering to differing standards and regulations regarding AWS use (Blanchard et al., 2024). Such disparities can exacerbate the overall risk magnitude, highlighting the need for harmonized international standards on AWS deployment.

In addition, regulatory inefficiencies, especially concerning liability mechanisms, can result in unclear distribution of primary (damage prevention) and secondary (damage retribution) costs related to AWS deployment. Another crucial extrinsic risk is political risks, such as geopolitical tensions and shifts in political climates, which may prompt AWS misuse. For instance, political pressure might accelerate the deployment of AWS without comprehensive testing across all potential conflict scenarios. Finally, economic factors, including financial instability, supply chain disruptions, economic sanctions, or challenges in military procurement, constitute extrinsic risks as far as they can impede technological advancements or the effective implementation of security measures for AWS. These economic challenges can stifle the development of safety mechanisms and reduce operational risks. For instance, financial constraints might limit research and development funding on AWS safety research, hindering the incorporation of advanced safety techniques, or evaluation and assurance methods. Similarly, supply chain issues could delay the delivery of critical components necessary for AWS reliability and security enhancements.

Ancillary risks emerge from risk regulation itself. These risks come in several types. For instance, innovation risk may arise if the regulatory framework is excessively restrictive, potentially impeding innovation and development. This is closely tied to opportunity risk, which refers to the potential loss of benefits that arise from advanced technologies. For instance, overregulation could inhibit the use of AWS, thereby preventing the realization of possible benefits, such as greater adherence to the cardinal principles of IHL. At the same time, as already seen, under-regulation also presents risks, e.g., unharmonized and unequal global regulation, or inconsistent enforcement of the same, for AWS. This is a significant concern given the cross-border use of these technologies. Lastly, a risk-based paradigm also entails its own ‘risks’ as strategic aims, values, and even alliances can be undermined by excessive focus on risk management (Beck, 1992).

Table 1 Risk matrix

***Likelihood*(%)**						
Severity	Major	L	H	H	E	E
	Serious	*N*	H	H	E	E
	Moderate	*N*	L	H	H	H
	Light	*N*	*N*	L	H	H
	Minimal	*N*	*N*	*N*	*N*	*N*
		0–0.20	0.20–0.40	0.40–0.60	0.60–080	0.80–1

Adapted from (Ni et al., 2010)

2.2 A Risk Matrix for AWS

The risk matrix is a semi-quantitative tool commonly used in risk analysis (Ni et al., 2010). However, it has been criticised, prompting recommendations for cautious application (Cox Jr, 2008; Markowski and Mannan 2008; Aven & Cox Jr, 2016). Despite this, it is an effective tool for selecting risk control measures when accurately refined.

Risk matrices typically integrate two main input variables — severity of harm and likelihood (frequency) of occurrence — to calculate a risk rating by positioning each severity-likelihood combination within the matrix. It segments the severity of consequences, probability, and the resultant risk magnitude into various levels, each described qualitatively and quantified on specific scales. In our methodology, these qualitative descriptions are informed by identifying and analysing the four main risk determinants, their drivers, and both extrinsic and ancillary risks, as detailed in Sect. 2.1.

To illustrate this, we have hypothesized categorizing the severity of harm into five levels – e.g., major, serious, moderate, light, and minimal. Similarly, the likelihood of occurrence can range from 0 (impossible) to 1 (certain).[12] The combination of severity and probability sets four (preliminary) risk tolerance thresholds that AWS can fall into, namely negligible (N), low (L), high (H), and extreme (E)[13]:

This risk matrix is a foundational tool for assessing the risk magnitudes associated with using AWS. [14] It should be regarded as an initial framework rather than a conclusive method for managing AWS risk, given that standard risk matrices have faced critique and calls for refinement. One approach to enhance the accuracy of risk calculations involves representing all variables within the matrix as fuzzy sets – which include linguistic terms of variables and description range – instead of fixed metrics, allowing for a more nuanced interpretation of risk factors (Markowski and Mannan 2008). An alternative method involves replacing the discrete categories in tradi-

[12] Generally, risk severity is influenced by all the four risk determinants, while the likelihood is mainly influenced by the interaction between hazard and response drivers.

[13] These tolerance thresholds are merely illustrative. However, by mirroring the risk levels outlined in the EU AI Act, they offer an example of how the AWS risk taxonomy might be structured.

[14] The four risk magnitude coefficients might also be compared to the four tolerance thresholds of the EU AIA (i.e., minimal, limited, high, and unacceptable risks). While the AIA does not regulate AWS, it provides a standard many AI models must adhere to. Given AI technologies are often dual-use, the EU AIA could thereby indirectly effect AWS or AI-driven safety components for AWS.

tional risk matrices with continuous scales for consequence and likelihood, leading to a continuous probability consequence diagram (Duijm, 2015). This method can offer advantages, although it does not fix all the limitations of standard risk matrices, including potential ambiguity in assessing consequences. It allows for the representation of uncertainty using uncertainty bands.[15]

Although this paper does not aim to determine the definitive risk matrix approach or the best way to overcome it, it emphasizes the critical need for ongoing research in this domain, particularly considering the opportunity to develop a risk-based regulatory framework for AWS.

A final critical aspect to consider is the subjective assignment of likelihood and consequence within risk matrix approaches (Duijm, 2015). This is a challenge, especially regarding risk-based regulations for AWS. Subjectivity can lead to divergent risk assessments for comparable AWS deployments across various regulatory entities or inconsistencies within the same organization across different periods. Such variability contributes to confusion among AWS developers and users, fostering uncertainty regarding compliance obligations and the sufficiency of implemented safety protocols. Ultimately, if risk assessments are perceived as arbitrary due to subjective risk factor evaluations, AWS developers and operators may question the regulatory framework's validity and applicability, undermining the rationale of a risk-based approach.

To address and mitigate the negative impacts of this subjectivity, it is crucial to use one of the above-mentioned three criteria for risk assessment: *background knowledge*. Acknowledging that risk magnitude can be determined through various methods, including qualitative approaches (e.g., the IPCC framework) and semi-quantitative methods (risk matrices), it becomes imperative to incorporate background knowledge. This knowledge outlines the depth and reliability of the information that underpins judgments concerning the likelihood and severity of a particular risk event (Aven, 2017). Its inclusion aims to clarify the subjectivity inherent in risk assessments, acknowledging that information volume, origin, and reliability shape the subjectivity observed in these evaluations. Making explicit the role of background knowledge in shaping risk assessments can help stabilize interpretations and enhance the transparency of risk-based regulatory frameworks for AWS.

Assessing background knowledge for AWS requires analysing the availability of relevant data, the soundness of assumptions, understanding of the subject matter — often judged by the known accuracy of predictive models used, expert consensus, and evaluating the base knowledge for unexpected events (Aven, 2017, p. 44; 2015, p. 26). This may involve examining data on AWS performance from simulations, field trials, and real-world deployments, as well as studying perceptions of military AWS use (Rosendorf et al., 2022). It is also essential to secure agreement among military experts, AI researchers, and ethicists on the ethical deployment and risks of AWS (Mitchell, 2019).

After gathering background knowledge, the risk matrix can be refined following Aven's recommendation: adjusting probability values based on the robustness of the

[15] Other alternative approaches recur to using bow ties and Bayesian influence diagrams (Meyer and Reniers 2022).

underlying knowledge, categorized as ‘strong’, ‘medium’, or ‘weak’. For example, if the probability of erroneous targeting by AWS (event 1) is initially assessed as low (0–0.20%) but the knowledge supporting this assessment is ‘weak’, its risk magnitude could be equivalent to that of an adversarial attack on AWS (event 2), which has a higher assessed probability (0.20–0.40%) but is supported by ‘strong’ knowledge (Aven, 2015). This approach ensures that the risk matrix reflects not just the likelihood of events, but the confidence in the data underpinning those likelihood assessments.

However, a mere calculus is insufficient to establish risk tolerance thresholds for policymaking. This is because risk measurements are somewhat abstract and can be influenced by the subjective inclinations of the concerned community. In Sect. 3, we shall outline a method for gauging this community’s risk tolerance and how it links with our previously defined risk thresholds. Before doing so, we provide a case scenario to demonstrate the practical application of our approach. We underscore that this case scenario is illustrative as it is up to relevant actors – e.g. states and/or international organizations – to determine risk thresholds.

2.3 Case Scenario: AWS Deployment in a Peacekeeping Mission

Consider a peacekeeping mission led by an international organization operating in a region experiencing armed clashes between insurgent groups and governmental forces. The mission’s mandate includes protecting civilian populations, monitoring ceasefire lines, and supporting humanitarian relief. The peacekeeping force is considering deploying AWS to improve perimeter security, detect infiltration attempts, and reduce harm to peacekeepers and civilians. How might a decision-maker gauge the risk level of deploying AWS in this complex setting? To do so, it is necessary to identify and understand the relevant risk factors and establish the components of the risk matrix, with particular attention to the severity levels.

The selected AWS can autonomously patrol designated areas, identify potential threats, and issue warnings. However, flaws in the AWS target recognition algorithm may cause it to misclassify non-threats as threats [H: Hazard]. This threat interacts with nearby populations [E: Exposure], where civilians are regularly present within the operational area of the AWS. Limited understanding by civilians that an AWS is operating nearby, and no clear means to signal noncombatant status elevate [V: Vulnerability], making it more likely that a misclassification results in serious harm. Let us assume that methods to reduce the probability and severity of false-positive misclassification [R: Response] – e.g., human confirmation before engagement – have been neglected in this scenario, e.g., due to deficient training.

In this scenario, the severity levels of the risk matrix could be qualified as follows.

Major. Catastrophic impact, including multiple civilian fatalities, large-scale destruction of critical infrastructure, severe and long-lasting psychological harm to the community, or substantial or severe violations of relevant international law.

Serious. Serious harm to one or more civilians (e.g., an injury requiring extensive medical attention), actions that deeply undermine trust in the mission without reaching the scale of mass casualties, widespread infrastructure collapse, or significant legal violations.

Moderate. Notable yet not life-threatening harm (e.g., minor injuries to civilians that can be treated locally, temporary damage to infrastructure, or short-term confusion/disruption among the population), combined with no sustained violation of relevant international law.

Light. Minimal physical or psychological harm (e.g., minor injuries, quickly resolved confusion or panic), negligible, easily repaired property damage, and no substantial breach of legal norms, though the incident may still raise questions about diligence and best practices.

Minimal. Practically no direct harm to persons or infrastructure (e.g., false alarms that cause momentary caution but no actual injury or damage), negligible operational impact, and no identifiable breach of IHL or related legal frameworks.

From a regulatory perspective, a level of risk from serious to major would generally discourage the deployment of AWS without additional safeguards. However, as discussed above, the ultimate decision for establishing risk thresholds in deploying AWS depends on the risk appetite of the relevant community. While some may choose to deploy AWS despite the high risk, others may not. The community's risk appetite can shape both the estimation of likelihood and the interpretation of severity thresholds. Therefore, to have a complete picture and implement this risk-based approach, we need to explore the role of risk appetite in these assessments.

3 Risk Appetite: International Humanitarian Law (IHL) and Just War Theory (JWT)

The risk magnitude coefficient, quantifiable through specific metrics, should be evaluated considering the concerned community's risk appetite. This involves adapting the assessment of potential consequences to reflect the community's values, objectives, and practices. For this reason, we suggest that a possible method to calculate risk appetite is by balancing the values and objectives of that community against the concrete importance, for that community, of deploying AWS, to maximize the marginal benefits between these two aspects. The result of this two-phase process could serve as the primary metric for establishing risk tolerance thresholds that would dictate varying levels of safeguards for AWS. These could range from complete prohibition to voluntary adherence to codes of conduct.

To gauge risk tolerance, we need to identify the target community and determine the risks it would deem acceptable, considering cultural habits, values, objectives, and socio-political conditions. In essence, what level of risk is the community willing to take on? Given the inherent complexity of cultures and political contingencies, particularly in heterogeneous and broad contexts, we shall consider them as variables to be determined contingently, and we will not factor them into this analysis. Instead, we shall concentrate on values and objectives.

Concerning the target community, multiple stakeholders, including military and defence entities, civilian populations, and humanitarian organizations, are relevant. Given AWS' cross-national deployment and the necessity for widespread stakeholder coordination, setting risk tolerance thresholds with the international community in mind is crucial. The risk tolerance of entities comprising this community is shaped by concerns over global stability, compliance with relevant international law (and case law) and human rights legislation, possible misuses of military technologies, and arms proliferation. The international community's risk tolerance is critical to establishing a global AWS risk assessment framework.[16]

Understanding this risk tolerance requires evaluating how using a specific AWS (i.e., a specific technical configuration within a particular context), with its associated risk magnitude, interferes with the community's normative background. This process requires ensuring that defence interests in deploying an AWS do not come at the cost of respecting fundamental values and objectives. The interference of an AWS with specific values or objectives can be judged as either broadly acceptable, acceptable under some conditions, or completely unacceptable.[17] In other words, the risk tolerance filters the quotient of risk magnitude through evaluative judgments and trade-offs, leading to different thresholds of risk tolerance for each specific AWS.[18]

The sources for discerning the international community's normative background may vary between deliberative fora, but are likely to include legislative texts comprising IHL, international human rights law (IHRL), and other relevant international laws, including non-proliferation and disarmament law.[19] We illustrate the methodology by considering IHL and the ethical framework of JWT to infer the guiding values and objectives. Other principles could be considered.

IHL and JWT share key principles concerning the legitimacy and usage of weapons, including distinction, proportionality, necessity, and precautions in attack.[20] Risk tolerance hinges on balancing these principles and the (social) interest in adopting AWS. This may be achieved by evaluating the degree to which the risk magnitude of an AWS interferes with, and potentially undermines, IHL and JWT principles. The interference should be counterbalanced by the concrete benefit expected from using the AWS. The process of weighing rights against public interest, or against other rights, is a standard practice in law-making and judicial reasoning (Bongiovanni and Valentini 2018). Indeed, legal principles are optimization commands, and our goal is to realize them to the greatest extent possible given the juridical and factual

[16] We highlight, therefore, that the establishment of a risk tolerance, and the implementation of a risk framework, will depend much for its success on the state of global multilateralism.

[17] Assuming only three risk tolerance thresholds, as in the ALARP ('As Low As Reasonably Possible') UK risk management approach (UKHSE, 2001).

[18] Political and normative factors influence risk magnitude, as assigning weight to hazard, exposure, vulnerability, and response drivers reflects risk perception. Thus, risk appetite is partly inferred from risk magnitude assessment. However, risk magnitude, while not fully objective, can mirror the broad intersubjectivity among experts (Aven et al., 2011) and is more objective than risk appetite, which heavily depends on subjective perceptions.

[19] International Humanitarian Rights Law (IHRL) may be relevant here, although its status and its applicability to AWS is contested with the United Nations Human Rights Council (UNHCR) having requested its Advisory Committee to prepare a study into this area (United Nations Human Rights Council, 2022; Blanchard, 2023a, b).

[20] Precautions in attack is most closely paralleled in JWT by 'due care'. Due care entails normative assumptions about the fair distribution of risks in war, for instance, the obligation of due care, which requires combatants to accept greater risks to themselves to ensure that they hit only the right target in order to diminish the risks to noncombatants (Walzer, 1977, p. 156; Orend, 2001, 12–13; Avishai Margalit and Michael Walzer, 2009; McMahan, 2010).

possibilities (Alexy, 2000). As Aharon Barak points out, the basic rule of balancing establishes: "[…] a general criterion for deciding between the marginal benefit to the public good and the marginal limit to human rights." Furthermore, that

> […] the extent that greater importance is attached to preventing the marginal limit to a human right and to the extent that the probability of the right being limited is higher, the marginal benefit to the public interest brought about by the limitation must be of greater importance, of greater urgency, and possessing a greater probability of materializing (Barak, 2010, p. 11).

To show how these types of balancing choices work, Alexy has provided a quantitative criterion, the well-known Weight Formula ($W_{x,\,y}$):

$$W_{x,y} = \frac{I_x \cdot W_x}{C_y \cdot W_y}$$

The Weight Formula can be applied as a tool for making explicit (or establishing) the international community's tolerance for the risk of AWS: I_x would correspond to the degree of interference an AWS has on a (set of) principle(s) within the scope of IHL or JWT in a concrete situation, e.g., the principle of distinction (P_x) as it is impacted using a drone with autonomy in its critical functions. The coefficient of the risk magnitude of the AWS will largely influence the coefficient of this factor. C_y would correspond to the concrete importance of satisfying the colliding interest brought about by the AWS, e.g., national security or strategic deterrence (P_y). Finally, W_x and W_y denote the abstract values the community assigns to the two principles.[21]

This formula puts into practice a proportionality test. For Alexy, it is a way to operationalize the concept of strict proportionality, which echoes the principle of proportionality required by IHL. In both instances, proportionality aims to evaluate whether the significance of fulfilling a particular interest or principle can justify infringing upon another.

The components of the Weight Formula can indeed be assigned numerical values using arithmetic or geometric sequences (such as 0, 2, 4, 16). Inherent in the structure of the formula, a higher result in the ratio tends to prioritize safeguarding the principle that is being impacted using the specific AWS. This reflects a low-risk tolerance threshold for that specific AWS. Conversely, lower values in the ratio suggest a greater willingness to sacrifice the principle – but never eliminate it – corresponding to a higher risk tolerance threshold.

In this article, we refrain from assigning numerical values or ranges to the principles of IHL or JWT. This decision stems from the absence of established regulatory frameworks or consensus on AWS' risk levels, coupled with the contentious nature of assigning numerical values, which can be perceived as arbitrary.[22] The Weight

[21] Often it is not possible to infer the abstract value of some principles, especially where they are incommensurable. This would mean that greater importance will be assigned to the concrete circumstances.

[22] The methodology of proportionality does not exclusively depend on numerical magnitudes, such as Alexy's formula, but can still engage with quantitative reasoning. An example includes using Pareto supe-

formula is a conceptual guide for how one might systematically balance normative elements, like legal principles and rights. Its goal is to inform the judgments that decision-makers — e.g. policymakers and legislators — routinely employ when reconciling multiple interests, by offering a structured way to consider how different values might be weighed against one another. However, it should be noted that the Weight formula is neither the only nor a universally applicable way to balance conflicting principles and rights.

In conclusion, our objective is to clarify the essential factors and their interactions, highlighting that effective risk-based regulation for AWS necessitates a combination of evaluating risk magnitude, background knowledge, and risk appetite. The latter is influenced by subjective preferences and becomes apparent in the trade-offs between principles and interests. We believe that the most effective approach, morally and legally, is to assess this risk appetite by finding a proportional balance between these principles and interests.

4 Conclusion

The international discourse on regulating and potentially using AWS in armed conflict can progress only when founded on a shared evaluation of the risks associated with AWS. To support this discussion, we have introduced a semi-quantitative approach, drawing on risk science literature, to assess and evaluate these risks and to establish risk tolerance thresholds for policymaking activities. If – or indeed *once* – a risk threshold for underpinning international regulations is determined, mechanisms for the monitoring and verification of compliance with such regulations will still be required. Article 36 of Additional Protocol 1 to the Geneva Conventions has been proposed as a potential mechanism for ensuring compliance in the design and development of AWS. However, the lack of its explicit and implicit content raises doubts about its (current) suitability for this purpose.[23] In either case, we hope the above risk assessment approach can help advance such efforts.

Author contributions Alexander Blanchard and Claudio Novelli are both first authors. All authors read and approved the final manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability Data sharing is not applicable: We do not analyse or generate any datasets, because our work proceeds within a theoretical approach.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest related to this paper.

riority criteria to assess proportionality, thus maintaining the foundational quantitative assessment principle (Sartor, 2013).

[23] See Farrant and Ford 2017; Boulanin, 2015; Chengeta, 2016; Meier, 2016; Copeland, Liivoja, and Sanders 2022; McFarland and Assaad 2023.

References

Abrahamsen, E., Bjorheim, H. B., Abrahamsen, M. F., Milazzo, & Jon Tømmerås, S. (2018). Using the ALARP principle for safety management in the energy production sector of chemical industry. *Reliability Engineering & System Safety*, *169*(January), 160–165. https://doi.org/10.1016/j.ress.2017.08.014

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access: Practical Innovations, Open Solutions*, *6*, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385

Alexy, R. (2000). On the structure of legal principles. *Ratio Juris*, *13*(3), 294–304. https://doi.org/10.1111/1467-9337.00157

Anderson, K., & Waxman, M. C. (2017). Debating autonomous weapon systems, their ethics, and their regulation under international law. SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=2978359

Anthony, T. & Cox Jr, L. (2008). What's wrong with risk matrices?? *Risk Analysis*, *28*(2), 497–512. https://doi.org/10.1111/j.1539-6924.2008.01030.x

Automated Decision Research. (2023). *Convergences in state positions on human control*. Automated Decision Research.

Aven, T. (2015). *Risk analysis* (2nd ed.). Wiley. https://doi.org/10.1002/9781119057819.ch2

Aven, T. (2017). Improving risk characterisations in practical situations by highlighting knowledge aspects, with applications to risk matrices. *Reliability Engineering & System Safety*, Special Section: Applications of Probabilistic Graphical Models in Dependability, Diagnosis and Prognosis, 167 (November):42–48. https://doi.org/10.1016/j.ress.2017.05.006

Aven, T., & Cox Jr, L. A. (2016). National and global risk studies: How can the field of risk analysis contribute?? *Risk Analysis*, *36*(2), 186–190. https://doi.org/10.1111/risa.12584

Aven, T., Renn, O., & Rosa, E. A. (2011). On the ontological status of the concept of risk. *Safety Science*, *49*(8), 1074–1079. https://doi.org/10.1016/j.ssci.2011.04.015

Barak, A. (2010). Proportionality and principled balancing. *Law & Ethics of Human Rights*, *4*(1), 1–16. https://doi.org/10.2202/1938-2545.1041

Baybutt, P. (2014). The ALARP principle in process safety. *Process Safety Progress*, *33*(1), 36–40. https://doi.org/10.1002/prs.11599

Beck, U. (1992). *Risk society: Towards a new modernity*. Sage.

Bhuta, N., & Stavros-Evdokimos Pantazopoulos (2016). Autonomy and Uncertainty: Increasingly Autonomous Weapons Systems and the International Legal Regulation of Risk. In *Autonomous Weapons Systes: Law, Ethics, Policy*, edited by Nehal Bhuta, Susanne Beck, Robin Geiβ, Hin-Yan Liu, and Claus Kreβ. Cambridge: Cambridge University Press.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, *84*(December), 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Blanchard, A. (2023a). Autonomous force beyond armed conflict. *Minds and Machines*, March. https://doi.org/10.1007/s11023-023-09627-z

Blanchard, A., & Taddeo, M. (2023b). Jus in Bello necessity, the requirement of minimal force, and autonomous weapons systems. *Journal of Military Ethics*, *0*(0), 1–18. https://doi.org/10.1080/15027570.2022.2157952

Blanchard, A. & Bruun, L. (2024). Bias in military artificial intelligence. Stockholm International Peace Research Institute. https://doi.org/10.55163/CJFT9557

Blanchard, A., & Taddeo, M. (2022). *Autonomous weapon systems and Jus ad bellum*. AI & SOCIETY. https://doi.org/10.1007/s00146-022-01425-y. March.

Blanchard, A., Thomas, C., & Taddeo, M. (2024). *Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance*. AI & SOCIETY. https://doi.org/10.1007/s00146-024-01866-7. February.

Bode, I., & Huelss, H. (2018). Autonomous weapons systems and changing norms in international relations. *Review of International Studies*, *44*(3), 393–413. https://doi.org/10.1017/S0260210517000614

Bongiovanni, G. & Valentini, C. (2018). Balancing, Proportionality and Constitutional Rights'. In *Handbook of Legal Reasoning and Argumentation*, edited by Giorgio Bongiovanni, Gerald Postema, Antonino Rotolo, Giovanni Sartor, Chiara Valentini, and Douglas Walton, 581–612. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-9452-0_20

Boulanin, V. (2015). Implementing Article 36 Weapon Reviews in the Light on Increasing Autonomy in Weapon Systems. Stockholm: Stockholm International Peace Research Institute. https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf

Boulanin, V., Carlsson, M. P., Goussac, N. & Davidson, D. (2020). Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control'. Stockholm International Peace Research Institute and the International Committee of the Red Cross. https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0

Bruun, L. (2024). Towards a Two-Tiered approach to regulation of autonomous weapon systems: Identifying pathways and possible elements. Stockholm International Peace Research Institute. https://doi.org/10.55163/LPED7967

Canellas, M. C., & Haga, R. A. (2015). Toward Meaningful Human Control of Autonomous Weapons Systems through Function Allocation. In *2015 IEEE International Symposium on Technology and Society (ISTAS)*, 1–7. https://doi.org/10.1109/ISTAS.2015.7439432

CCW GGE (2019). Guiding principles affirmed by Th group of governmental experts on emerging technologies in the area of lethal autonomous weapons system (Annex III). In, *CCW/MSP/2019/9*. https://www.ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf.Geneva, SwitzerlandUnited Nations Office of Disarmament Affairs.

Chandler, K. (2021). Does Military AI Have Gender? Understanding Bias And Promoting Ethical Approaches In Military Applications of AI. United Nations Institute for Disarmament Research. https://doi.org/10.37559/GEN/2021/04

Chengeta, T. (2016). Are autonomous weapons systems the subject of Article 36 of additional protocol I to the Geneva conventions. *UC Davis J Int'l L & Pol'y*, *23*, 65.

Conway, M. (2020). Smashing the Patriarchy The Feminist Case Against Killer Robots'. Centre for Feminist Foreign Policy. https://static1.squarespace.com/static/57cd7cd9d482e9784e4ccc34/t/5f356f1e5eb59d07e78fb329/1597337376556/Smashing+the+Patriarchy_+The+Feminist+Case+Against+Killer+Robots.pdf

Copeland, D., Liivoja, R., & Sanders, L. (2022). The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems. *Journal of Conflict and Security Law*, November. https://doi.org/10.1093/jcsl/krac035

Duijm, N. J. (2015). Recommendations on the use and design of risk matrices. *Safety Science*, *76*(July), 21–31. https://doi.org/10.1016/j.ssci.2015.02.014

Edney-Browne, A. (2019). The psychosocial effects of drone violence: Social isolation, Self-Objectification, and depoliticization. *Political Psychology*, *40*(6), 1341–1356. https://doi.org/10.1111/pops.12629

Etzioni, A. (2018). Pros and Cons of Autonomous Weapons Systems (with Oren Etzioni). In *Happiness Is the Wrong Metric: A Liberal Communitarian Response to Populism*, edited by Amitai Etzioni, 253–63. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-69623-2_16

Farrant, J., & Christopher, M. F. (2017). Autonomous weapons and weapon reviews: The UK second international weapon review forum. *International Law Studies*, *93*, 35.

Figueroa, M., Díaz, A. H., Orozco, J., Martínez, & Wanda Muñoz, J. (2023). The risks of autonomous weapons: An analysis centred on the rights of persons with disabilities. *International Review of the Red Cross*, *105*(922), 278–305. https://doi.org/10.1017/S1816383122000881

Flage, R., & Aven, T. (2015). Emerging Risk – Conceptual definition and a relation to black Swan type of events. *Reliability Engineering & System Safety*, *144*(December), 61–67. https://doi.org/10.1016/j.ress.2015.07.008

Geiss, R. (2016). Autonomous Weapons Systems: Risk Management and State Responsibility. Geneva.

Holland Michel, A. (2020). The Black Box, Unlocked: Predictability and Understandability in Military AI. United Nations Institute for Disarmament Research. https://doi.org/10.37559/SecTec/20/AI1
Horowitz, M. C. (2021). When speed kills: Lethal autonomous weapon systems, deterrence and stability. *Emerging technologies and international stability*. Routledge.
Hua, S. S. (2019). Machine learning weapons and international humanitarian law: Rethinking meaningful human control. *Geo J Int'l L*, *51*, 117.
ICRC. (2021). *ICRC position on autonomous weapon systems & background paper*. International Committee of the Red Cross.
ICRC (2022). Autonomous Weapons: The ICRC Remains Confident That States Will Adopt New Rules. Statement. International Committee of the Red Cross. 11 March 2022. https://www.icrc.org/en/document/icrc-autonomous-adopt-new-rules
Johnson, J. (2020). Artificial intelligence, drone swarming and escalation risks in future warfare. *The RUSI Journal*, *165*(2), 26–36. https://doi.org/10.1080/03071847.2020.1752026
Laird, B. (2020). The risks of autonomous weapons systems for crisis stability and conflict escalation in future U.S.-Russia Confrontations. RAND Corporation. 3 June 2020. https://www.rand.org/blog/2020/06/the-risks-of-autonomous-weapons-systems-for-crisis.html
Leys, N. (2018). Autonomous weapon systems and international crises. *Strategic Studies Quarterly*, *12*(1), 48–73.
Longpre, S., Storm, M., & Shah, R. (2022). Lethal Autonomous Weapons Systems & Artificial intelligence: Trends, Challenges, and Policies. Edited by Kevin McDermott. *MIT Science Policy Review*, *3* (August):47–56. https://doi.org/10.38105/spr.360apm5typ
Margalit, A. & Walzer, M. (2009). *Israel: Civilians & combatants*. The New York Review of Books.
Markowski, A. S., & Sam Mannan, M. (2008). Fuzzy Risk Matrix. *Journal of Hazardous Materials*, Papers Presented at the 2006 Annual Symposium of the Mary Kay O'Connor Process Safety Center, 159(1):152–57. https://doi.org/10.1016/j.jhazmat.2008.03.055
McFarland, T. (2022). Minimum levels of human intervention in autonomous attacks. *Journal of Conflict and Security Law*, *27*(3), 387–409. https://doi.org/10.1093/jcsl/krac021
McFarland, T., & Assaad, Z. (2023). Legal reviews of in situ learning in autonomous weapons. *Ethics and Information Technology*, *25*(1), 9. https://doi.org/10.1007/s10676-023-09688-9
McMahan, J. (2010). The just distribution of harm between combatants and noncombatants: The just distribution of harm between combatants and noncombatants. *Philosophy & Public Affairs*, *38*(4), 342–379. https://doi.org/10.1111/j.1088-4963.2010.01196.x
Meier, M. W. (2016). Lethal autonomous weapons systems (LAWS): Conducting A comprehensive weapons review. *Temp Int'l & Comp LJ* 30.
Melzer, N. (2008). The Principle of Distinction under International Humanitarian Law'. In *Targeted Killing in International Law*, edited by Nils Melzer, 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199533169.003.0011
Meyer, T., & Reniers, G. (2022). Engineering risk management. In *Engineering Risk Management*. De Gruyter. https://doi.org/10.1515/9783110665338
Mitchell, C. (2019). When laws govern laws: A review of the 2018 discussions of the group of governmental experts on the implementation and regulation of lethal autonomous weapons systems student notes. *Santa Clara High Technology Law Journal*, *36*(4), 407–432.
Molyneux, B. S. & Camilla. (2021). The Human Cost of Remote Warfare in Yemen'. In *Remote Warfare: Interdisciplinary Perspectives*. E-International Relations. https://www.e-ir.info/2021/02/16/the-human-cost-of-remote-warfare-in-yemen/
Moore, C. (1990). Unpredictability and undecidability in dynamical systems. *Physical Review Letters*, *64*(20), 2354–2357.
Muñoz, W., Mariana, & Díaz (2021). The Risks of Autonomous Weapons: An Intersectional Analysis. SEHLAC. https://img1.wsimg.com/blobby/go/98c6dc90-096f-4389-9309-f1a33c0cad73/downloads/SEHLAC%20Autonomous%20Weapons%20and%20Interseccionalit.pdf?ver=1689286516267
Musiolik, T. H., & Cheok, A. D. (Eds.). (2021). *Analyzing future applications of AI, sensors, and robotics in society: Advances in computational intelligence and robotics*. IGI Global. https://doi.org/10.4018/978-1-7998-3499-1
Ni, H., Chen, A., & Chen, N. (2010). Some extensions on risk matrix approach. *Safety Science*, *48*(10), 1269–1278. https://doi.org/10.1016/j.ssci.2010.04.005
NIST. (2023). *AI risk management framework (AI RMF 1.0)'. NIST AI 100-1*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). AI risk assessment: A Scenario-Based, proportional methodology for the AI act. *Digital Society*, *3*(1), 13. https://doi.org/10.1007/s44206-024-00095-1

Nurick, L. (1945). The distinction between combatant and noncombatant in the law of war. *American Journal of International Law*, *39*(4), 680–697. https://doi.org/10.2307/2193409

Orend, B. (2001). Just and lawful conduct in war: Reflections on Michael Walzer. *Law and Philosophy*, *20*(1), 1–30. https://doi.org/10.2307/3505049

Peel, J. (Ed.). (2010). Global risk governance and its legitimacy. *Science and risk regulation in international law* (pp. 12–57). Cambridge University Press. Cambridge Studies in International and Comparative Lawhttps://doi.org/10.1017/CBO9780511779879.003

Quinn, G. (2021). Report of the special rapporteur on the rights of persons with disabilities. *UN Doc A*, *76*, 146. https://documents.un.org/doc/undoc/gen/n21/196/98/pdf/n2119698.pdf?token=Zv9opyjmTJwPDE7rnf&fe=true

Ramsay-Jones, H. (2019). Racism and Fully Autonomous Weapons. Soka Gakkai International. https://www.ohchr.org/sites/default/files/Documents/Issues/Racism/SR/Call/campaigntostopkillerrobots.pdf

Rice, H. G. (1956). On completely recursively enumerable classes and their key arrays. *Journal of Symbolic Logic*, *21*(3), 304–308. https://doi.org/10.2307/2269105

Roff, H. M. & Moyes, R. (2016). Meaningful Human Control, Artificial Intelligence and Autonomous Weapons. In *Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons, Geneva, Switzerland*.

Roff, H. M., & Danks, D. (2018). Trust but verify: The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics*, *17*(1), 2–20. https://doi.org/10.1080/15027570.2018.1481907

Rosendorf, O., Smetana, M., & Vranka, M. (2022). Autonomous weapons and ethical judgments: Experimental evidence on attitudes toward the military use of killer robots. *Peace and Conflict: Journal of Peace Psychology*, *28*(2), 177–183. https://doi.org/10.1037/pac0000601

Samuel, A. L. (1960). Some moral and technical consequences of Automation–A refutation. *Science*, *132*(3429), 741–742. https://doi.org/10.1126/science.132.3429.741

Sartor, G. (2013). The logic of proportionality: Reasoning with Non-Numerical magnitudes. *German Law Journal*, *14*(8), 1419–1456. https://doi.org/10.1017/S2071832200002339

Sassoli, M. (2014). Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified. *International Law Studies* 90 (1). https://digital-commons.usnwc.edu/ils/vol90/iss1/1

Scharre, P. (2016). *Autonomous weapons and operational risk*. Center for a New American Security.

Schwarz, E. (2021). Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control. *Philosophical Journal of Conflict and Violence*. https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/74360/Schwarz%20Autonomous%20Weapons%20Systems,%20Artificial%20Intelligence,%20and%20the%20Problem%20of%20Meaningful%20Human%20Control%202021%20Accepted.pdf?sequence=2

Simpson, N. P., Katharine, J., Mach, A., Constable, J., Hess, R., Hogarth, M., Howden, J., Lawrence, et al. (2021). A framework for complex climate change risk assessment. *One Earth*, *4*(4), 489–501. https://doi.org/10.1016/j.oneear.2021.03.005

Taddeo, M. (2024). *The ethics of artificial intelligence in defence*. Oxford University Press.

Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics*, *28*(5), 37. https://doi.org/10.1007/s11948-022-00392-3

Taddeo, M., Ziosi, M., Tsamados, A., Gilli, L., & Kurapati, S. (2022). *Artificial intelligence for National security: The predictability problem*. Centre for Emerging Technology and Security.

Thurnher, J. S. (2018). Feasible Precautions in Attack and Autonomous Weapons. In *Dehumanization of Warfare: Legal Implications of New Weapon Technologies*, edited by Wolff Heintschel von Heinegg, Robert Frau, and Tassilo Singer, 99–117. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-67266-3_6

UKHSE (2001). Risk Management: Expert Guidance - ALARP at a Glance. 2001. https://www.hse.gov.uk/managing/theory/alarpglance.htm

UN Women (2023). Women Are Increasingly At-Risk in Conflict, Underrepresented in Peace Processes, According to UN Secretary-General Report. UN Women – Headquarters. 24 October 2023. https://www.unwomen.org/en/news-stories/feature-story/2023/10/women-are-increasingly-at-risk-in-conflict-underrepresented-in-peace-processes-according-to-un-secretary-general-report

United Nations Human Rights Council (2022). 51/22 Human Rights Implications of New and Emerging Technologies in the Military Domain. Resolution Adopted by the Human Rights Council on 7 October 2022. A/HRC/RES/51/22. New York: United Nations General Assembly. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G22/520/64/PDF/G2252064.pdf?OpenElement

van den Boogaard, J. (2024). Warning! Obstacles Ahead! The Regulation of Autonomous Weapons Systems in the GGE LAWS. *Opinio Juris* (blog). 4 March 2024. https://opiniojuris.org/2024/03/04/warning-obstacles-ahead-the-regulation-of-autonomous-weapons-systems-in-the-gge-laws/

Walzer, M. (1977). *Just and unjust wars: A moral argument with historical illustrations*. Basic Books.

Weiss, K., Khoshgoftaar, T. M., DingDing, & Wang (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9. https://doi.org/10.1186/s40537-016-0043-6

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, *131*(3410), 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Winter, E. (2022). The compatibility of autonomous weapons with the principles of international humanitarian law. *Journal of Conflict and Security Law*, *27*(1), 1–20. https://doi.org/10.1093/jcsl/krac001