03

ASSER INSTITUTE
Centre for International & European Law

RESEARCH PAPER

Iterative Assessment for Military Artificial Intelligence (AI) Systems

Jonathan Kwik

June 2025

asser.nl

This text may be downloaded for personal research purposes only. Any additional reproduction for other purposes, whether in hard copy or electronically, requires the consent of the author. If cited or quoted, reference should be made to the full name of the author, the title, the working paper or other series, the year, and the publisher.

© Jonathan Kwik, 2025
Forthcoming in: Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds.), *Decision at the Edge: Interdisciplinary Dilemmas in Military Artificial Intelligence*, The Hague: T.M.C. Asser Press (forthcoming 2025)
SSRN Asser page
www.asser.nl
Cite as: ASSER research paper (2025 - 03)
Contact author: j.kwik@asser.nl; j.h.c.kwik@gmail.com

Abstract

Artificial intelligence (AI) introduces unprecedented uncertainty in military operations. This is particularly evident in AI-enabled autonomous weapon systems (AWS) and decision support systems (DSS), which not only influence critical battlefield decisions but also pose novel and unpredictable risks. While some risks can be anticipated and managed *ex ante*, many remain inherent and unavoidable, given the complex, dynamic, and adversarial nature of the environments in which these systems operate. Even AI operators acting in good faith may face situations in which unforeseeable civilian harm occurs, despite rigorous review and careful deployment. In practice, many such incidents will be characterised as 'accidents'—a reality of war that International Humanitarian Law is expected to tolerate.

This paper challenges that assumption, arguing that even *a priori* unpredictable AI failures can be mitigated—if not prevented—through an iterative approach. By systematically integrating insights from post-deployment assessments, this approach enables decision makers to update their understanding of edge cases and other 'known unknowns' that emerge during real-world use, providing essential insights to inform future AI deployment. It proposes an *Iterative Assessment* framework—implemented through two complementary mechanisms: *Iterative Review* and *Iterative Assessment in Deployment*. This framework represents best practice for managing uncertainty and minimising civilian harm in the use of military AI. While initial accidents may be unavoidable, their recurrence can be significantly reduced through a structured iterative process of reporting, analysis, and adaptation. Those committed to the responsible use of military AI should embed this framework as a core component of operational planning and legal compliance.

Keywords

Uncertainty, Artificial Intelligence (AI), Accident, Autonomous Weapons Systems, Decision Support Systems, Predictability, Weapons Review, International Humanitarian Law

Table of contents

1 Introduction

As artificial intelligence (AI) becomes increasingly embedded in military operations, its users will be confronted with heightened uncertainty, unpredictable risks, and system vulnerabilities. Many of these challenges will only reveal themselves *ex post*—after failures have already resulted in humanitarian harm, such as civilian injury or loss of life. In practice, such incidents are often characterised as unavoidable 'accidents', and tolerated as the inevitable by-products of war. This paper challenges that assumption and proposes an *Iterative Assessment* framework as a best practice approach for militaries seeking to proactively mitigate AI-related civilian harm in dynamic operational environments.

The framework rests on two foundational tenets. On one hand, it acknowledges that failures involving military AI may be unavoidable at the outset, and that decision makers cannot be expected to know the unknowable. On the other, it argues that the overarching spirit of International Humanitarian Law (IHL) imposes a duty on belligerents to take all feasible steps to prevent the recurrence of such harm once relevant risks become knowable. To this end, the Iterative Assessment framework introduces a two-tiered mechanism designed to support rapid, adaptive mitigation of emerging AI-related risks.

The paper proceeds as follows. Sect. 2 contextualises the discussion by examining how the heightened unpredictability of AI systems undermines the effectiveness of traditional quality control and risk mitigation mechanisms in preventing repeated failures. Building on this, Sect. 3 provides a semi-technical overview of the various sources of AI uncertainty, demonstrating that many stem from immutable properties of the technology and, therefore, cannot be eliminated through technical means alone—nor can they be fully addressed through *ex ante* precautions. To address this challenge, the paper advocates for the adoption of an iterative mindset—one that operationalises IHL obligations continuously over time, rather than at isolated junctures. With this philosophy in mind, Sects. 4 and 5 explore the application of Iterative Assessment in the review and targeting phases, respectively. On this basis, Sect. 6 then synthesises the model into a visual operational guide to support practical implementation. Finally, Sect. 7 concludes with key reflections and actionable recommendations for embedding Iterative Assessment into military practice.

Before proceeding, it is helpful to clarify the nature and purpose of the Iterative Assessment framework. The term 'framework' is used deliberately, in preference to alternatives such as 'norm' or 'principle'. This reflects the fact that Iterative Assessment is not intended to represent a binding legal obligation *lex lata*, particularly under more restrictive interpretations of IHL. Rather, it refers to a structured set of measures and policies that decision makers can adopt to minimise risk to civilians during the development, testing, and deployment of AI systems. These measures are rooted in a core philosophy of iteration and continuous improvement. At the same time, Iterative Assessment is neither radical nor entirely novel. On the contrary, many of the practices recommended in this paper already exist within contemporary military doctrine and have been advanced in scholarship.[1] The primary challenge lies not in their conceptual novelty, but in their consistent implementation—something that in practice is often constrained by operational or logistical realities.

As to what Iterative Assessment can achieve, its limits must also be recognised. It cannot eliminate all uncertainty: this has always been an inherent feature of military operations. However, when properly implemented, Iterative Assessment can strengthen

[1] Additionally, under more permissive interpretations of IHL, one could argue that Iterative Assessment aligns with the 'constant care' obligation set out in Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), opened for signature 8 June 1977, 1125 UNTS 3, entered into force 7 December 1978 (API), art. 57(1).

adherence to IHL by helping to prevent avoidable civilian harm. In essence, it aims to reduce the number of instances where harm occurs not due to recklessness, but because decision makers lacked access to information that was, in *theory*, knowable. While not a panacea, Iterative Assessment offers a meaningful and practical standard for those seeking to responsibly integrate AI into military operations.

Finally, this paper proceeds on the assumption that AI users act in good faith, seeking to uphold IHL and to manage uncertainty responsibly. It presumes that all pre-existing obligations related to testing, review, and precautions in attack, as required under IHL, have been duly fulfilled. It does not address bad-faith scenarios, such as the deliberate deployment of 'black box' systems without any attempt to understand their functions.[2] Iterative Assessment is designed to complement—not replace—these foundational IHL obligations. The objective of this paper is to demonstrate that even when conventional IHL requirements are properly met, there remains scope for further improvement through the adoption of an iterative approach.

[2] Beyond constituting bad faith, many commentators argue that deploying AI systems under such conditions constitutes a clear violation of IHL. For example, see Sullivan and Ricket 2024; Holland Michel 2020; Kwik and Van Engers 2021.

2 The Case for Iterative Assessment

Warfare has always been inherently fraught with uncertainties. Legal decisions concerning the use of force, and the manner in which it is applied, often rely on 'determinations of fact that may be difficult to make'.[3] Scientific advancements, while falling short of offering full transparency on the battlefield,[4] have nonetheless helped to reduce this uncertainty. Modern militaries increasingly rely on more precise weaponry, advanced sensors, satellite imagery, and other tools to peer through the proverbial fog of war and support legal compliance.[5] Against this backdrop, AI is often presented as the next technological leap: a system capable of processing vast quantities of signal data, enhancing intelligence gathering, and improving situational awareness.[6]

Despite these potential advantages, the integration of AI into military operations risks *reducing* transparency, owing to the unique properties of the technology itself. AI systems exhibit unpredictable characteristics—including the black box phenomenon, unintuitive failure modes, continuous learning, and emergent, unprogrammed behaviours—all of which introduce novel uncertainties and pose serious challenges for normative legal compliance.[7] The unpredictability of AI has long been recognised as a formidable obstacle to the application of IHL.[8] Among the concerns raised, it has been submitted that such technological uncertainties may:

Complicate exhaustive testing, validation, and quality assurance processes, as AI systems are likely to encounter real-world deployment conditions that could not have been anticipated during development;[9]

Undermine weapons reviews, as required under Additional Protocol I (API) to the Geneva Conventions (GCs) or implied in customary IHL,[10] due to difficulties in fully understanding how an AI system will behave in operational environments;[11]

Invalidate IHL compliance assessments made at the moment of activation,[12] as unanticipated behaviours may render such evaluations obsolete immediately after deployment; and

Impair the proper implementation of precautionary obligations, as AI users may be unable to reliably predict the system's behaviour or anticipate its operational consequences.[13]

These concerns are not limited to AI deployed in use-of-force systems, such as autonomous weapons systems (AWS), but are equally relevant to AI-enabled decision-support systems (DSS).[14] Moreover, they span across the entire lifecycle of an AI system—from development through deployment to post-deployment—complicating efforts to ensure continued compliance with IHL across each phase.[15]

[3] US Department of Defense 2015, Section 5.3.1.
[4] Stewart 2011, p. 293.
[5] Ekelhof 2018, p. 76; Haque 2012, p. 110; Schmitt and Schauss 2019, p. 152.
[6] Mikhailov 2023, p. 2.
[7] These factors are comprehensively discussed in Sect. 3.
[8] For example, see ICRC 2016, pp. 2-3.
[9] Lohn 2020, p. 1.
[10] API, n.1; Switzerland 2016, para. 23; Klonowska 2022.
[11] Holland Michel 2020, p. 1
[12] ICRC 2019, p. 3.
[13] Liu 2016, p. 335.
[14] In this paper, AWS refers to physical weapon systems equipped with AI components that assist in use-of-force functions (e.g., autonomous target recognition and engagement). In contrast, DSS refers to AI-powered software designed to assist human operators in making use-of-force decisions (e.g., target identification, target sorting, and course-of-action recommendations). For a detailed distinction between the two, see ICRC and Geneva Academy 2024, pp. 9-10.
[15] An exception to this observation arises where reviews are deemed applicable solely to AWS and not to DSS, as there is ongoing debate over whether DSS qualify as 'means and methods of warfare', thereby triggering the legal duty to review. See also, Sect. 4.

Various approaches have been proposed to respond to this unpredictability. Measures such as explainability requirements, designed to combat the black box nature of modern algorithms (as implemented by the United States (US) and the North Atlantic Treaty Organisation (NATO)),[16] enhanced modelling-based testing and review processes,[17] and prohibitions on online learning post-deployment,[18] have all been proposed as potential safeguards.[19] Best practice for ensuring compliance with IHL would undoubtedly involve the adoption of such measures and standards by militaries, where feasible.[20] Still, one may question whether certain sources of AI uncertainty can ever truly be eliminated.

As one example, the military environment is inherently complex and dynamic, presenting an almost infinite range of possible input-output pairings. As such, 'edge cases and unforeseen performance-degrading elements always remain a possibility',[21] regardless of how diligently systems are reviewed, or how well an operator claims to understand both the system and the operational environment.[22] Even the most genuine and good-faith belligerents will inevitably encounter surprises.

Consider these hypothetical situations, which highlight key challenges that will be referenced throughout this discussion to illustrate the importance of Iterative Assessment for AI systems:

> ***Drone (AWS):*** During a one-month operation, autonomous drones mistakenly targeted and killed several local civilians after traditional clothing patterns worn in the region triggered AI hallucinations.[23] Despite extensive review and testing of the drone model—including assessments accounting for typical regional attire—this edge case went undetected. The hallucination was only activated when the patterns were viewed from a specific angle, a factor not identified during testing and evaluation. Similar incidents were reported across multiple platoons operating the drones in the same region.
>
> ***Targeting Adviser (DSS)***: A DSS powered by a large language model (LLM) with a chat interface was deployed to assist officers in target analysis. After one month of use, a scathing NGO report revealed that numerous approved targets had, in fact, been civilian. Although the system had performed reliably during prior testing and deployments, officers—under operational pressure—had unknowingly framed their chat prompts with urgency. The LLM misinterpreted this as a directive to prioritise speed over accuracy, leading to target misidentifications.[24] Even the system's engineers were shocked to discover this emergent behaviour, i.e., that the LLM had learnt to infer urgency from the prompt structure and had dynamically altered its outputs accordingly.

Despite the seriousness of these outcomes, it is highly probable that such harms would be dismissed as '[m]istakes or faulty weaponry'—inevitable occurrences in the conduct of war.[25]

Militaries have, at times, been criticised for too readily characterising such incidents as 'accidents', when, in reality, they may stem from systemic flaws in targeting practices (e.g., reliance on poor or outdated intelligence or perceived behavioural patterns).[26] However, the *Drone* and *Targeting Adviser* cases truly appear to constitute genuine accidents: rigorous testing and evaluation were undertaken, and there is no indication that the operators acted recklessly (or even negligently).[27] In the *Drone* case, none of the platoon

[16] US Department of Defense 2023, p. 4; NATO 2021a, para. C.
[17] Meier 2019, p. 30
[18] That is, where the AI system is permitted to continue adapting its weights based on inputs obtained from the operational environment after deployment; see Defense Innovation Board 2019, p. 46.
[19] For example, Moyes 2019, p. 11; see also Boothby 2018, p. 151.
[20] IHL 'accounts for the limited and unreliable nature of information in armed conflict'; see US Department of Defense 2015, Section 5.4.3.2. However, a minimum epistemic threshold is required for decision makers to make reasoned and legally compliant choices, and reasonable commanders are required to take active steps to reduce uncertainty where feasible; see Thorne 2020, p. 2; Schmitt and Schauss 2019, p. 152; As Kalmanovitz remarks, '[w]hen there is uncertainty, the relevant practical question to ask is what steps have been taken to sufficiently determine and limit the risks created'; Kalmanovitz 2016, p. 156.
[21] Kwik 2024a, p. 70.
[22] Many other sources of AI-induced uncertainty present challenges that lack straightforward solutions, as discussed in Sect. 3.
[23] Hallucinations occur 'when an AI model generates incorrect or misleading information but presents it as if it were a fact'. For several examples, see Guinness 2024.
[24] Research has shown that prompt formulation can influence the output of large models such as LLMs in unforeseen ways; see Sharma et al. 2023.
[25] Bartels 2013, p. 280.
[26] See Hathaway and Khan 2024, pp. 40-56.
[27] Kwik 2024a, p. 367, defines a genuine accident as one in which 'notwithstanding all feasible precautions employed, a failure occurs; … [or] [w]here there was no foreseeability of the outcome, despite due diligence or extensive technical training of persons involved'.

commanders were likely aware of the specific edge case, nor had the issuing authority identified it, despite extensive review and testing. While basic precautions in attack would have accounted for encounters with local civilians wearing traditional clothing,[28] available system data gave no indication that this would trigger fatal misclassifications. Similarly, in the *Targeting Adviser* case, the officers had no reason to suspect that the LLM could infer urgency from their prompts and dynamically adjust its outputs in a harmful way.[29] Nor had this behaviour been detected during the system's vetting process by the issuing authority.[30] At both the review and targeting stages, relevant IHL obligations appear to have been fulfilled to the extent required by law. With the constraints of the available information, one could argue that nothing more could reasonably have been done.

This paper adopts a different perspective, arguing that these two cases nonetheless reveal opportunities for significant improvements—measures that could have substantially reduced their humanitarian impact. The main critique lies not in the initial failures themselves, but in the actions (or inactions) taken *during* and *after* their occurrence. While many IHL safeguards focus on *ex ante* measures—those undertaken prior to the adoption of a system or the launch of an attack—there is a clear need for additional mechanisms to protect civilians from AI-related failures and vulnerabilities that only emerge, and are only knowable, *ex post*.

In response, this paper proposes Iterative Assessment as a unifying framework to manage the persistent uncertainties inherent in military AI. The core of the concept lies in its first word, *iteration*: the process of repeatedly refining decisions or actions based on accumulated experience, new information, and evolving understanding, with the aim of continuous improvement.[31] This approach departs slightly from conventional risk mitigation methods (e.g., implementing a fixed explainability standard, or disabling online learning), which often treat uncertainty as a static problem. By contrast, Iterative Assessment embraces uncertainty as dynamic and evolving. It calls on decision makers to adopt a longitudinal mindset: drawing lessons from past deployment iterations, proactively monitoring ongoing operations, and remaining responsive to emerging insights over time.

The remainder of this paper explores how the various forms of Iterative Assessment can be applied to AI systems. It demonstrates how its adaptive, forward-looking framework can enhance IHL compliance by reducing AI-related uncertainties to more manageable and acceptable levels.[32]

[28] Under IHL, commanders are required to collect such necessary information to implement precautions and minimise risks to civilians; see Oeter 2020, p. 354.
[29] It is theorised that as models grow larger, they will not only develop this capacity but also face selection pressure to predict users' preferences and conditions, adapting their outputs accordingly—potentially in undesirable ways. See Ngo et al. 2022, pp. 3-4.
[30] Since testing aims to replicate real-world conditions as closely as possible, the issuing authority might have identified this niche LLM behaviour by simulating a high-stress operational environment, prompting testers to write overly emotive inputs. However, this would not have guaranteed the behaviour's detection, as testers may have responded differently from the operators in the actual incident above—whether due to personality differences or simply their awareness of being in a simulated scenario.
[31] This philosophy will be familiar to readers acquainted with Bayes' theorem, and rightly so. The approach proposed in this paper is strongly inspired by Bayesian reasoning, which is particularly well-suited to situations where reasoned predictions must be made under conditions of partial information. As McGrayne notes, it is 'an evolving system, which each new bit of information pushe[s] closer and closer to certitude'; see McGrayne 2011, p. 8. To maintain focus on the doctrinal and practical elements of this paper, further references to Bayes' theorem will not be made. However, those familiar with Bayesian theory will recognise its influence throughout this discussion.
[32] IHL tolerates decision-making conditions short of absolute certainty, provided such decisions fall within the margins of reasonableness. See Kwik 2024a, p. 70.. As Henderson also notes, '[i]t cannot be expected that armed conflict will be reduced to the point where a commander can act only when he or she is 100 percent certain in all cases': Henderson 2009, p. 164.

3 Sources of Uncertainty in Military AI

This section outlines key factors that exacerbate uncertainty in the deployment and operational use of AI systems in military contexts. While not exhaustive, the discussion highlights how many of these factors stem from properties unique to AI technologies. As such, they introduce novel challenges that not only compound existing sources of operational uncertainty, but also give rise to entirely new forms of it—shaped by the dynamic and opaque nature of modern AI systems.[33]

3.1 Algorithmic and Environmental Uncertainty: A Toxic Mix

Modern AI—particularly machine learning (ML) models,[34] such as deep neural networks (DNN)—offers solutions to computational problems that are intractable with rule-based approaches alone.[35] These models now dominate the field of AI,[36] enabling capabilities such as sensor fusion and complex target recognition analysis,[37] that would otherwise be unachievable. However, unlike rule-based approaches, ML models are inherently less interpretable: their behaviour is not based on predefined rules,[38] but instead emerges from input-output pairings shaped by patterns and inferences derived from training data.[39]

Due to the difficulty of discerning their internal logic, modern AI models are often described as ***opaque* or *black boxes***.[40] Such models 'do not allow cognitive access to how they have arrived at a particular output, or what input factors or a combination of input factors have contributed to the decision-making process or outcome'.[41] Nevertheless, not all ML models are *necessarily* black boxes. Some architectures are designed to permit external understanding of their decision-making logic. These models can be transparent by design,[42] or they may rely on added explanatory mechanisms intended to describe their internal processes (even where the underlying model itself remains opaque).[43] However, it remains doubtful whether such measures can fully resolve the problem. Not all models can be transparent by design, and explanatory mechanisms may themselves be incorrect or misleading.[44] As a result, uncertainty—though varying in degree—persists. While human operators can assess AI model performance through empirical testing and approximate its decision-making logic,[45] it remains impossible to predict how an ML model will behave across all possible input-output pairings.

Indeterminacy in AI arises not only from the difficulty of interpreting a system's internal logic. It also stems from the ***variability of the input space***. The more dynamic and complex the environment in which an AI system operates, the greater the range of potential inputs—leading to increased behavioural variance and making exhaustive testing or prediction increasingly difficult.[46] One method of mitigating this variability is to structure the operational environment in ways that constrain the range of inputs an AI system may encounter. For example, in the case of autonomous vehicles, efforts to enhance predictability have included implementing

[33] For example, determining the legality of potential targets inherently involves a degree of uncertainty, even when employing sophisticated identification methodologies; see Corn 2012, p. 43
[34] ML enables AI to 'learn without being explicitly programmed' and 'involves the use of algorithms to parse data and learn from it, and making a determination or prediction as a result'; see Roy 2018, p. 14.
[35] Russell and Norvig 2021, p. 651.
[36] Oniani et al. 2023, p. 3.
[37] Cranny-Evans 2024; Meng et al. 2022, p. 2084.
[38] Deng 2015, p. 25.
[39] Abaimov and Martellini 2020, p. 14.
[40] Molnar describes a black box mode as one 'that cannot be understood by looking at their parameters': Molnar 2019, p. 13.
[41] Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659) 2020, p. 4
[42] See Barredo Arrieta et al. 2020, p. 88.
[43] Rosenfeld 2021, p. 2.
[44] See Silva et al. 2023, p. 3; Van der Waa et al. 2021, p. 2.
[45] See Kwik 2024a, pp. 79-91.
[46] Russell et al. 2015, p. 108.

clear signage and reducing environmental clutter—measures that limit the scope for unforeseen interactions.[47] Structuring the environment in this way narrows the system's potential input space and, and in turn, reduces uncertainty.

In contrast, battlefields are rarely amenable to such structuring.[48] The military environment is dynamic, complex and adversarial.[49] It is continuously evolving, stochastic, and filled with external variables beyond the AI user's control—including hostile agents actively seeking to evade or sabotage operations.[50] Granted, this volatility is not new: the battlefield has always been messy and unpredictable.[51] However, when combined with the input-output variability inherent in modern AI systems, the result is an explosion of possible outcomes that renders uncertainty virtually unavoidable.

Unpredictability is further exacerbated by the ***unintuitive nature of AI failure triggers***.[52] ML models often exhibit 'counterintuitive or poorly described failure modes that do not follow historical patterns of human, or even software, failures', making them difficult to anticipate based on test distributions alone.[53] AI systems may perform consistently over extended periods, appearing predictable and understandable, only to react abruptly and inexplicably to seemingly innocuous inputs. For example, white paint may cause an autonomous car to misinterpret its surroundings and crash,[54] or grey texture patches may lead an AI classifier to misidentify an elephant as a cat.[55] Identifying and addressing all such edge cases is a formidable challenge, as they often arise from highly specific combinations of inputs that evaluation datasets cannot comprehensively capture.[56] Consequently, sooner or later, deployed AI systems will inevitably encounter unforeseen edge cases, and in worst case scenarios, may fail catastrophically—as illustrated by the *Drone* incident.

This problem is further compounded by the ***adversarial*** nature of military environments. Unlike civilian AI applications, it should be assumed that 'enemies will likely attempt to exploit vulnerabilities of the system'.[57] While some adversaries may target conventional vulnerabilities (e.g., such as jamming communications or disrupting data feeds),[58] others will specifically seek to exploit the unintuitive failure triggers described above to increase their chances of success. One example of such a Counter-AI (CAI) measure is the so-called input or evasion attack, where adversaries create 'malicious input data designed to deceive AI algorithms, leading to incorrect predictions or classifications'.[59] These attacks exploit edge cases within the target AI system—vulnerabilities that may not even be known to the system's operators.[60] While vulnerability to such adversarial techniques can be mitigated through robustness training and other defensive measures,[61] complete security is impossible.[62] As a result, even where all other factors are known and foreseeable, adversarial interference may still induce unpredictable AI behaviour.[63]

Online learning, whereby a model 'uses continuous cycles of retraining and model updating from new data input',[64] further aggravates uncertainty. While this capability offers advantages in dynamic environments—particularly against adaptive adversaries on the battlefield—it also allows AI systems to 'update the model hypothesis constantly' in response to changing conditions.[65] Online learning can serve as an effective

[47] For example, this approach has been implemented in controlled environments, such as factory settings or on roadways; see Boulanin 2016, p. 18.
[48] Certain environments, such as underwater domains, are inherently less complex or cluttered. Structuring a military environment (e.g., by removing elements that could trigger unpredictable behaviour or by adding markers to provide contextual clues for AI) is likely only feasible in areas under near-complete friendly control; see Kwik 2024a, p. 208.
[49] Kwik 2024a, pp. 70-71.
[50] Tolk 2015, p. 298. See generally Russell and Norvig 2021, p. 925.
[51] This has even been normatively acknowledged by IHL through provisions incorporating qualifications such as 'may be expected' and 'reasonableness'. Herbach 2012, p .17.
[52] See generally Kwik 2024a, Ch. 5.
[53] Gilmer et al. 2018, p. 6.
[54] In this incident, a Tesla vehicle failed to detect 'a large white 18-wheel truck and trailer crossing the highway' due to the contrast between the truck's white colour and a bright spring sky: see Yadron and Tynan 2016.
[55] See Geirhos et al. 201 In this experiment, researchers overlaid a grey texture image onto a tabby cat, causing AI classifiers to misidentify it as an elephant. In contrast, human observers still recognised the distinctive shape and contours of a cat.
[56] Hendrycks et al. 2020, p. 1. See also Flournoy et al. 2020, p. 8.
[57] Scharre 2016, p. 1. As a NATO document highlights, '[s]ome state and non-state actors will likely seek to exploit defects or limitations within our AI technologies': see NATO 2021a, para. 14.
[58] Such attacks may include jamming, spoofing, electromagnetic pulse (EMP) attacks, denial-of-service attacks, etc: see Wilson 2020.
[59] Mikhailov 2023, p. 3.
[60] One might question why an adversary would be aware of an exploitable edge case while the AI user remains unaware. This discrepancy arises from the asymmetry between attackers versus defenders in such scenarios. An attacker succeeds by identifying and exploiting *any* possible edge case through an input attack, whereas a defender can only prevent or counter such attacks if they have prior knowledge of *that specific* vulnerability. Therefore, an AI user's lack of awareness of an exploited edge case does not necessarily indicate negligence regarding system security. For further discussion on this dynamic, see Kwik 2024a, p. 151.
[61] For example, see Mikhailov 2023, pp. 3-4.
[62] Gilmer et al. 2018, p 2.
[63] For a complete taxonomy on how adversaries can employ anti-AI attacks to sabotage or even take control over military AI systems, see Kwik 2024b, pp. 5-11.
[64] Nelson et al. 2014, p. 1,
[65] Das 2021.

countermeasure against adversarial attacks. For example, spam filters are designed to continually adapt to evolving spamming techniques.[66] Given these benefits, online learning may be invaluable for many military AI applications.[67] However, this flexibility comes at a cost: it undermines predictability. It becomes increasingly difficult to anticipate *when* and *how* the AI model will adapt its algorithm to environmental changes. As Leslie highlights, 'the unbounded complexity of the world makes anticipating all of its pitfalls and detrimental variables veritably impossible'.[68]

Finally, looking ahead, researchers have raised concerns regarding the potential for ***emergent behaviour*** to become a significant challenge for military AI users.[69] As AI systems become increasingly complex, they may exhibit 'actions or patterns that weren't explicitly programmed … but developed as a natural outcome of its complexity and interactions'.[70] In military contexts, emergent behaviour is likely to manifest in two (non-mutually exclusive) ways. First, goal-oriented multi-agent systems, such as autonomous swarms,[71] may be programmed with simple and linear instructions at the unit level yet behave unpredictably as a collective. These interactions can introduce stochastic and nonlinear dynamics, making it difficult to predict how the system will function under real-world conditions.[72] Second, as model size, dataset complexity, and training duration increase, AI systems—particularly LLMs—have exhibited 'sharp and unpredictable changes in model outputs'.[73] Emergent behaviour is a double-edged sword: it 'presents opportunities, but also poses important risks'.[74] LLMs may unexpectedly develop highly beneficial capabilities, such as zero-shot learning,[75] but may equally exhibit undesirable behaviours, such as reinforcing a user's preconceived views through an emergent ability to psychoanalyse user input.[76] The *Targeting Adviser* case exemplifies how such emergent behaviour can lead to harmful consequences in practice.[77] In particular, because such behaviour does not follow linear cause-and-effect reasoning,[78] it is exceedingly difficult to predict *what* will occur, *when*, and *why*.[79] In some circumstances, the only viable recourse may be to *recognise* the emergence of such behaviour and assess the *context* in which it arose.

3.2 Reflections on Uncertainty

At this juncture, several general observations can be made regarding the persistence of uncertainty in military AI and the challenges of navigating it. First, while some sources of uncertainty can be addressed to a certain extent, through technical or operational measures, they cannot be fully eliminated. For example, one could choose to 'freeze' an AI model upon deployment or enforce minimum explainability standards to mitigate uncertainty caused by online learning or opacity.[80] Other sources of uncertainty, however, are inherent to the use of modern AI in military contexts. Many stem from the fundamental nature of AI itself—including ML, opacity, and unintuitive failure triggers—or from the operational environment, characterised by highly variable input spaces, dynamic conditions, and adversarial interference. Beyond the *a priori* decision not to deploy the AI system at all,[81] these variables cannot be removed—they must be accepted.

Second, a good-faith AI user can nonetheless take steps to curtail uncertainty and manage variance.[82] Rigorous testing and validation across diverse conditions and

[66] King et al. 2020, pp. 96-97.
[67] Schuller 2017, p. 397.
[68] Leslie 2019, p. 34.
[69] For example, see Ekelhof and Paoli 2020; Boothby 2018, pp. 140-42.
[70] Gunnell 2023.
[71] These systems have become increasingly prevalent, especially in intelligence, surveillance, and reconnaissance (ISR) operations. However, they can also be employed more offensively, such as to overwhelm enemy air defence systems; see Government Accountability Office 2023; Safi 201
[72] Navarro and Matía 2013, p. 7.
[73] Schaeffer et al. 2023, p. 2 (emphasis removed).
[74] Steinhardt 2022.
[75] This phenomenon describes an AI system successfully performing a task it has never encountered before, despite seemingly lacking prior machine-learned experience in that domain; see Dickson 2022.
[76] See Perez et al. 2022, p. 10.
[77] The particular emergent behaviour exhibited by the DSS in the *Targeting Adviser* case is called 'sycophancy', see Kwik 2025a.
[78] For a general theoretical discussion on this phenomenon, see Fromm 2005.
[79] Ganguli et al. 2022, p. 15.
[80] These measures are not without trade-offs in terms of opportunity costs. For example, freezing AI models eliminates variance caused by continuous model evolution, simplifying testing and improving predictability: see ICRC 2018, para. 4 However, this rigidity reduces adaptability to changing circumstances and adaptive opponents. Similarly, while transparent models decrease uncertainty, they may also compromise efficiency, capability, and versatility, requiring additional time and resources for development: see Adadi and Berrada 2018, p. 5214.
[81] In general terms, the importance of this measure cannot be overstated. Indeed, the decision not to authorise an AI system (e.g., such as where its employment is deemed excessively risky) remains the most effective form of risk mitigation intervention at both the review and deployment stages, see Van den Boogaard and Roorda 2021 p. 433; Kwik 2024a, p. 35. This paper, however, focuses on scenarios where a reasonable decision maker proceeds with authorisation based on the system's apparent reliability and the available information at the time.
[82] This paper adopts a presumption based on this, as outlined in Sect. 1.

environments,[83] combined with extensive operator training, can improve awareness of AI vulnerabilities and failure states.[84] These measures can help reduce the risk of being blindsided by unexpected events, but they cannot eliminate all potential failures. Furthermore, because military AI systems operate in high-risk environments, even low-probability failures can lead to significant (humanitarian) costs, as demonstrated by the *Drone* and *Targeting Adviser* cases.[85]

The Iterative Assessment approach does not seek to eliminate these challenges, nor does it assume that all accidents are avoidable. At the same time, it is not defeatist. Rather, it is premised on the view that 'we can and should be able to recognise recurring patterns of misbehaviour, and to learn enough from past experience to be able to avoid or repair many of the common patterns'.[86] To implement this effectively, action is required at both the pre-deployment and operational stages, which I now explore.

[83] See, for example, Cherry and Johnson 2020, p. 13.
[84] Puscas 2023, p. 36.
[85] Bommasani et al. 2021, p. 116.
[86] Mogul 2005, p. 18.

4 Iterative Reviews

The review obligation is arguably the most critical guardrail against IHL violations at the pre-deployment stage.[87] Under its API formulation, Article 36 requires Parties to assess whether a new weapon, means, or method of warfare, would violate their international obligations, particularly those under IHL.[88] While the customary IHL duty to conduct such reviews is often considered less extensive than the API standard, there is growing consensus that the basic duty to ensure capabilities can be employed in conformity with IHL forms part of customary international law.[89] For the purposes of this analysis, it is assumed that reviews should extend to DSS, such as the system used in the *Targeting Adviser* example.[90] Beyond this assumption, this paper does not engage further with the doctrinal question,[91] and instead focuses on the *timing* of when such reviews are traditionally conducted.

4.1 Limitations of Conventional Reviews

In its most basic form, a weapon review is often seen as a discrete, one-time requirement performed during the weapon acquisition phase, typically just prior to adoption.[92] Given the generally low level of implementation among States,[93] it would be unsurprising if most AI-related reviews are similarly conducted in this manner in practice.[94] However, a slightly broader interpretation can be drawn from the API formulation, which requires reviews to be performed '[i]n the *study*, *development*, *acquisition* or *adoption* of a new weapon, means or method of warfare'.[95] Commentators favouring this approach consider that a narrow 'one-off' review is insufficient; instead, they advocate for reviews at 'each of the key decision points',[96] that is, 'at each stage of development and acquisition'.[97] This interpretation is reflected in the practice of many States,[98] supported by soft law instruments such as the Tallinn Manual,[99] and has been argued to be both more practical and economically efficient.[100] In addition to reviews at fixed milestones, some States also conduct *ad hoc* reviews in response to new pertinent information arising during development,[101] or when 'substantive changes occur'.[102]

One problem identified in the literature is that even this broader reading of the API review obligation may be insufficient to fully account for the unique properties of military AI, given that the 'development' of an AI system may continue beyond its point of adoption.[103] Unlike traditional weapons, an AI system can remain in a state of flux post-deployment. For example, if a DNN continues adjusting its weights—the parameters controlling the strength of connections between neurons—through frequent model updates or operational inputs (online learning) after deployment, this could render the legal review 'invalid immediately upon the use of the system'.[104] In itself, the possibility of adapting or altering already-adopted means or

[87] Sandoz et al. 1987, para. 1475.
[88] API, above n. 1, art. 36.
[89] See the Tallinn Manual 2.0 on the International Law Applicable to Cyber Warfare (Tallinn Manual 2.0), Schmitt 2017, p. 465; ICRC 2006, p. 4; Cf. Jevglevskaja 2018, p. 187.
[90] See Klonowska 2022.
[91] Even if one adopts the position that DSS fall outside the scope of the legal review obligation, the discussion in this Section remains relevant for AI weapon systems, such as those featured in the *Drone* case.
[92] Copeland 2014, p. 47.
[93] See ICRC 2006, p. 5; Boothby 2018, p. 18.
[94] Few States provide transparency regarding their review processes, making it difficult to ascertain how frequently such reviews are conducted before adoption: see Goussac 201
[95] API, above n. 1, art. 36 (emphasis added).
[96] McClelland 2003, p. 402.
[97] Fry 2006, p. 481; see also ICRC 2006, p. 23.
[98] For example, the UK Ministry of Defence 2016 conducts reviews 'at key milestones in the procurement process to assure the legality of a new weapon, means or method of warfare throughout its study, development, acquisition and adoption'; and the Belgian Armed Forces 2018, p. 7 conduct reviews '*[l]orsque la Défense étudie, met au point ou souhaite acquérir ou adopter une nouvelle arme, un nouveau moyen ou une nouvelle méthode de guerre*' [when Defence studies, develops, or intends to acquire or adopt a new weapon, a new means or a new method of warfare].
[99] Tallinn Manual 2.0 2017, p. 465.
[100] Conducting reviews at an 'early' stage allows for swift corrections in the design process or, the abandonment of a project entirely if it becomes evident that the final product would violate IHL and thus be unusable under any circumstances; see Daoust et al. 2002, p. 351; Boothby 2016, Sect. 15.2.
[101] For example, Belgian Armed Forces 2018 p. 8, '*Dans le cas où de nouvelles informations pertinentes sont rendues disponibles après le traitement du dossier par la CEJ, l'arme, le moyen ou la méthode de guerre sera à nouveau soumis à l'évaluation*' [If new pertinent information is made available after the completion of the review by the CEJ, the weapon, means or method of warfare shall be subjected to re-evaluation].
[102] Parks 2005, p. 134.
[103] See, for example, Meier 2019, p. 308; Boulanin et al. 2020, p. 13.
[104] Boulanin et al. 2020, p. 13. While this concern is often overstated—since modifying individual weights does not fundamentally alter an algorithm's behaviour—the underlying apprehension remains valid.

methods of warfare is not unique to AI systems. Some States and commentators have taken the position that a (re-)review would be required when a capability has 'undergone modification' or has been 'subject of an upgrade or other amendment that changes its combat performance'.[105] However, a particular challenge for software-based systems such as AI is that changes may manifest intangibly and incrementally (e.g., gradual shifts in weights over time).[106] Therefore, even if one accepts the position that '[a]ny changes to the system's operating state … would require the system to go through testing and evaluation again',[107] the precise timing of *when* this duty is triggered remains ambiguous.[108]

Most re-review clauses are triggered by modifications or alterations to a system. However, in both the *Drone* and *Targeting Adviser* cases, no factual changes had occurred within the systems themselves. The drones' algorithm had *always* contained the hidden edge case that caused hallucinations when viewing the clothing patterns from a specific angle. Similarly, the targeting adviser's LLM had *always* been capable of detecting user urgency and adapting its outputs accordingly. As a result, most re-review clauses would not have been triggered by these incidents. The main issue is that both re-review clauses and certain proposed solutions, such as 'freezing' the AI model,[109] focus on *factual transformations*—instances where the system or its operational environment factually undergoes an observable change after deployment.[110] However, in the *Drone* and *Targeting Adviser* cases, neither the system nor the environment changed; rather it was our *perception* and *understanding* that shifted. We became aware of a system behaviour that had previously been unknowable.[111] Let us call this an *epistemic transformation*.

When reviewing the various sources of AI uncertainty discussed in Sect. 3, it becomes clear that many factors stem more from epistemic limitations than from factual transformations. The latter is primarily driven by online learning and adaptive adversaries, whereas the former arises from structural challenges such as infinite input variance, algorithmic opacity, the unintuitive nature of AI failures, and emergent behaviour. Some have proposed modifications to review procedures to account for this epistemic gap,[112] such as placing greater emphasis on predictive modelling and simulations.[113] Nevertheless, it remains dubious whether the epistemic problem can ever be fully resolved, even with substantial reforms to how reviews are conducted. As one commenter notes, '[t]here are simply too many possible states and combination of states to be able to exhaustively test each one, and understanding where the boundary conditions are will be difficult'.[114]

4.2 The Iterative Approach

In response to the challenges outlined above, two possible courses of action can be taken. The first is to simply acknowledge these limitations and accept that incidents such as the *Drone* and *Targeting Adviser* cases may occur. Such incidents could be treated as accidents, attributable to *specific* environmental circumstances rather than fundamental flaws in the legal review process. Arguably, this is the default approach. For instance, a State would not ordinarily reassess the legality of its rocket system following an unprecedented malfunction caused by an unpredictable environmental factor, such as a flash-freezing or atmospheric turbulence incident prior to deployment.

The second, alternative approach is to adopt an *Iterative Review* process for AI systems. While uncommon in

[105] Boothby 2016, Sect. 15.2. Australia requires re-reviews for 'adaptations and modifications of existing weapons', and the UK conducts re-reviews when 'there is any change in a systems' use or capability'; Australia 2018 para. 9; UK Ministry of Defence 2016; see also ICRC 2006, n. 21.

[106] Similar discussions arose with respect to cyberweapons. See, for example, Tallinn Manual 2.0 2017, p. 466; Wallace 2018, p. 21.

[107] Sayler 2020.

[108] Boulanin et al. 2020, p. 13.

[109] Goussac 201

[110] Performance drift in AI systems can also result from environmental shifts (e.g., seasonal changes or adaptive adversaries), see Kwik 2024a, pp. 115-118. More comprehensive re-review clauses, such as those introduced in a recent DoD Directive, are therefore designed to trigger in response to *operational* changes. It reads:

> An autonomous weapon system that is a variant of an existing weapon system previously approved through this review will not be covered by previous approval if changes to the system *algorithms*, intended *mission set*, intended *operational environments*, intended *target sets*, or expected *adversarial countermeasures* substantially differ from those applicable to the previously approved weapon system so as to fall outside the scope of what was previously approved in the senior review. Such systems will require a new senior review before their formal development and again before fielding.

See US Department of Defense 2023, p. 15 (emphasis added).

[111] While these properties may be theoretically knowable, they were *factually* unknowable in the given circumstances due to constraints outlined in Sect. 3—particularly the challenge of testing for all possible edge cases and behaviours that emerge only under very specific conditions.

[112] For a broader discussion on AI certification, see Bakirtzis et al. 2022.

[113] Meier 2019, p. 309; Undersecretary of Defense for Acquisition, Technology, and Logistics 2012, p. 63.

[114] US Air Force Office of the Chief Scientist 2015, p. 23.

military contexts, this method is widely employed in the commercial software sector and in credentialing schemes (e.g., drug certification).[115] In contrast to conventional review methods, Iterative Review is characterised by the following attributes:[116]

> ***Provisional:*** The review process does not claim to be exhaustive or final in its findings. It accepts the possible existence of hidden unknowns within the system that could lead to high-cost failures in operational settings. Statements of (il)legality reflect only the reasonable belief of the reviewer based on information reasonably available at the time of assessment.
>
> ***Longitudinal:*** The review process does not have a fixed cut-off point but extends beyond adoption into the post-deployment phase.
>
> ***Adaptive*:** Previous assessments remain open to revision and refinement. As new data from deployment becomes available, legal assessments are updated accordingly. Necessary revisions must be actionable immediately, without undue procedural delay.
>
> ***Proactive:*** Rather than waiting for incident reports or violation complaints, data from deployments is proactively collected and integrated into the review process.[117]

These features are designed to specifically address review situations in which comprehensive and exhaustive testing (or anything that comes close to it) is impossible—a key challenge for military AI systems.[118] The *longitudinal* dimension is particularly critical, as rare incidents—such as the pattern recognition failure in the *Drone* case or the emergent user-interaction issue in the *Targeting Adviser* case—could only be identified through deployment. Yet, such discoveries may have profound legal implications. For example, a reasonable response to the *Targeting Adviser* case would be to revoke the DSS's prior legal approval, as its emergent ability to infer operator urgency from chat prompts introduces an unacceptable risk of recurrence under similar conditions.[119] To be effective, Iterative Reviews must also be *adaptive*: capable of rapidly revising prior legal assessments in response to new information and ensuring that any necessary restrictions or amendments take immediate effect.

A fundamental consequence of this approach is that no assessment can ever be truly definitive. Iterative Review acknowledges that unseen gaps in knowledge will always exist at the time of review—gaps that may later reveal a system to be fundamentally unlawful under IHL. At its core lies 'exploratory testing, shaped by insights from deployment'.[120] This *provisional* nature may sit uneasy with some, as it implies the deployment of military systems without a guarantee that their 'employment would, in some or all circumstances, be prohibited [under IHL]'.[121] This tension is unlikely to be fully resolved. However, it is important to recognise that conceptual disagreement already exists regarding the extent to which IHL compliance must be guaranteed in 'all circumstances'.[122] A factor in favour of permitting the iterative approach is that many interpret the Article 36 review obligation as requiring legality to only be assessed 'by reference to its normal expected use at the time the evaluation is conducted'.[123] It is submitted that it would be consistent with the spirit of the law if, at the time of review, the assessor is convinced to a reasonable degree—based on information reasonably available—that the system does not violate IHL, and issues a positive assessment, pending future information that may defeat this presumption.

At present, Iterative Review remains a recommendation *de lege ferenda*, especially given the issue of timing. It is difficult to derive a continued obligation for post-adoption reviews from the wording of API—'[i]n the study, development, acquisition or adoption'—let alone

[115] Meier 2019, p. 310; Bakirtzis et al. 2022, p. 1.
[116] The process described here follows the framework introduced by Bakirtzis et al. 2022.
[117] Cf. Hathaway and Khan 2024, p. 34.
[118] Trusilo 2023, p. 11.
[119] One could argue that this risk might be mitigated through alternative operational measures, such as instructing DSS users to avoid conveying urgency in their chat prompts. However, the core issue lies in the system's emergent ability to *infer operators' moods*, which raises the possibility of it detecting *other* emotional states and producing further unpredictable behaviours. This consideration also influences what constitutes a reasonable course of action in response to such findings at the operational level; see Sect. 5.
[120] Bakirtzis et al. 2022, p. 1.
[121] API, above n. 1, art. 36
[122] For an extended discussion, see Fry 2006, pp. 471-473, 501-503.
[123] Tallinn Manual 2.0 2017, p. 466. See also Bothe et al. 2013, p. 231; US Department of Defense 2015, Sect. 6.6.3.4.

from the more limited customary rule.[124] Moreover, Iterative Review is significantly more demanding than conventional review mechanisms. For it to be effective, a State would need to establish proactive and continuous monitoring systems, systematically collect deployment data, update models accordingly, and issue new reviews as required. These requirements may place a significant burden on States, particularly those with limited resources.[125] Nevertheless, States capable of implementing these mechanisms are urged to do so, as Iterative Review offers the most effective means of addressing epistemic transformations at the structural level.

At this point, one may ask how an Iterative Review mechanism would have responded to the *Drone* case. Given the rarity of the edge case, it is unlikely that a re-review would have found the drones structurally incapable of being used in accordance with IHL. However, the obligation of constant care would still necessitate the adoption of risk mitigation measures to protect civilians.[126] One possible review-level intervention could have been to modify the assessment to include operational restrictions,[127] such as prohibiting the drones' deployment in regions where traditional attire with distinct clothing patterns is prevalent.[128]

Despite the adaptive nature of Iterative Review, it is foreseeable that such an institutional-level change would require time to become fully effective—likely longer than the one-month period in which the *Drone* and *Targeting Adviser* incidents unfolded. During this period, the issue would have remained unaddressed, and civilians would have continued to face significant risk. Thus, while Iterative Review provides a valuable mechanism for long-term oversight, it is unlikely to respond rapidly enough to epistemic transformations in real time. For this reason, Iterative Review must be complemented by *Iterative Assessment in Deployment*, as discussed below.

[124] API, n.1, art. 36.
[125] States may alleviate this logistical burden by requiring manufacturers to incorporate monitoring capabilities into system designs, thereby facilitating oversight; see also below Sect. 7.
[126] Jenks and Liivoja 2018. See also Jensen 2021, p. 190.
[127] Assessors issuing reviews may 'attach conditions or comments … to be integrated into the rules of engagement or operating procedures': see ICRC 2006, p. 15.
[128] For an example of how such instructions can be integrated into military instructions or directives, see Kwik et al. 2025, pp. 18-21.

5 Iterative Assessment in Deployment

At the operational level, Iterative Assessment comprises two distinct yet equally important tasks: *Iterative Awareness* and *Proactive Response*.

5.1 Iterative Awareness

Iterative Awareness constitutes the fact-finding component of Iterative Assessment in Deployment. To achieve Iterative Awareness, AI operators leverage intelligence, surveillance, and reconnaissance (ISR) assets to monitor the civilian impact of a deployed system. Compared to standard contexts in which ISR is employed, Iterative Awareness differs slightly in its *timing*, emphasising *ex post* rather than *ex ante* data collection and analysis.

Commanders routinely collect and analyse intelligence prior to each attack—a necessary step for implementing precautionary measures and maintaining effective command-and-control (C2).[129] Much of the literature on military AI emphasises the importance of operators maintaining a thorough understanding of their AI systems, the battlefield, and adversary activity to ensure legal compliance and effective C2.[130] However, comparatively less attention has been given to the need for continuous monitoring of deployment outcomes,[131] which is equally crucial. *Ex post* awareness—the ability to assess the effects of a system after deployment—enables AI users to 'understand and assess the causes of malfunctions or undesired results from (normal) operation and to take appropriate action (technical and accountability) to prevent similar mistakes'.[132] As with pre-deployment legal reviews, there is a risk that the obligation to assess the lawfulness of an AI system is perceived as a one-time requirement, rather than a continuous duty persisting beyond initial deployment.

Conceptually, Iterative Awareness is already reflected in the established operational planning and targeting practices of many States. For example, the NATO targeting cycle consists of six phases,[133] with the final phase being 'Assessment'. This phase encapsulates the essence of remaining aware of past operations and using insights gained from previous deployments to inform future military actions:

> Assessment measures the extent to which the desired effects, regardless of the actions taken, have been created and recommends the extent to which further actions are required. It encompasses a physical, functional and system assessment. Assessment also contributes to wider operational and campaign assessment.[134]

The concept of *continuity*, central to the Iterative Assessment approach,[135] is also embedded within NATO and US military doctrine: '[c]ommanders and the staff must build and foster a comprehensive understanding of the operating environment and promote this understanding continuously throughout the entire operations planning process'.[136] Indeed, the notion of learning from prior iterations is widely recognised as a fundamental element of operational art. Doctrinal concepts such as 'lessons learned'—defined as 'the act of learning from experience to achieve improvements'—rely on field observations to refine

[129] Rosén 2014, p. 127. See also Australia 2006, para. 5.54; Schmitt 2010, para. 16.07.3.
[130] For example, Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS) 2021, para. 2a; Holland Michel 2020, p. 15. See also more generally NATO 2019a, pp. 2-4: 'A comprehensive analysis of the operating environment, its components, actors and their relationships is the beginning of the operations planning process'.
[131] Kwik 2022, p. 14.
[132] Van den Boogaard and Roorda 2021, p. 433.
[133] See NATO 2021b, para. 1.5.
[134] NATO 2021b, para. 1.5.1.f.
[135] This concept was similarly identified as essential in the discussion on iterative reviews; see Sect. 4.
[136] NATO 2019a, pp. 2-4. See also Curtis E. Lemay Center 2019, p. 79: 'Assessment is a continuous process that measures the overall effectiveness of employing joint force capabilities during military operations'.

military decision making.[137] After-action analysis is actively encouraged in order to enhance operational efficiency,[138] adapt to environmental changes,[139] and improve collateral damage estimation mechanisms.[140] At its core, this process enables adaptation:

> The outputs of an assessment communicate the effectiveness of the operation plan toward desired end states, describe risks involved in the accomplishment of the plan, and recommend necessary changes to the plan to attain a desired end state.[141]

Adaptation, however, was precisely what was lacking in both the *Drone* and *Targeting Adviser* cases. Despite the presence of a known edge case in the region (*Drone*) and an emergent vulnerability (*Targeting Adviser*), both systems remained in use for one month. This failure occurred because insufficient efforts were made to *detect* the mistaken targeting of civilians and to investigate *why* these errors were happening. These shortcomings were likely the result of practical and operational constraints. In the *Drone* case, misclassifications occurred across multiple platoons and throughout the region. Individual component commanders, operating in isolation, may have simply dismissed these incidents as accidental anomalies rather than recognising them as part of a broader pattern.[142] In the *Targeting Adviser* case, the failure to detect the high misclassification rate was likely driven by the same operational pressures that led the officers to frame chat prompts with urgency in the first place. Prioritising the execution of further attacks likely took precedence over verifying whether previous strikes had correctly targeted legitimate objectives.[143]

This highlights a crucial caveat to the claim that Iterative Awareness is an established practice: in reality, it often remains more *theoretical* than practical. While extensively discussed in military doctrine, its implementation in practice is questionable. Hathaway and Khan, for example, have identified several shortcomings in how the US conducts after-action assessments.[144] Among other issues, insufficient resources are dedicated to after-action data collection, and the data that is gathered is often incomplete or inadequate.[145] Additionally, the process is frequently slow and resource intensive, particularly in hard-to-access areas.[146] Even when after-action data is collected, it rarely accounts for *civilian* impact. Instead, most after-action assessments focus primarily on determining whether the *military* objective was achieved, such as assessing whether target re-engagement is necessary.[147] For Iterative Awareness to function effectively, it would therefore require the integration of a robust civilian harm tracking mechanism—one that ensures civilian impact is not treated as an afterthought but embedded as a core component of post-strike analyses.[148]

Another critical component of Iterative Awareness is *communication* across both horizontal and vertical axes. A strong civilian harm tracking mechanism could have alerted the *Targeting Adviser* operators to systematic errors in their DSS, prompting immediate corrective action. In contrast, in the *Drone* case, individual component commanders would likely have been unable to detect a structural issue in the region based on a single misclassification report. Had they received consolidated reports from multiple platoons indicating recurring misidentification of civilians wearing regionally distinctive attire, they would have been more likely to recognise that the drones were systematically encountering a persistent edge case linked to the local civilian population.

It should be noted that operational- or tactical-level conclusions will often be less precise, particularly in complex deployment environments. In the *Drone* case, for example, it is unlikely that field reports would have

[137] NATO 2019b, para. E.1; LEX-7.
[138] NATO 2021b, pp. 1-4.
[139] US Joint Chiefs of Staff 2018, pp. VI-1.
[140] Schmitt et al. 2017, p. 306.
[141] US Joint Chiefs of Staff 2018, pp. VI-2.
[142] Isolated weapon or software failures occur regularly across various military capabilities and do not necessarily prompt commanders to question the reliability or lawfulness of a system; see Scharre 2016, pp. 25, 38.
[143] A similar dynamic is reported to have allegedly occurred during the Israeli Defence Forces' use of the Lavender system in Gaza following the October 2023 attack by Hamas; see Abraham 2024.
[144] Hathaway and Khan 2024, pp. 32-6
[145] Ibid.
[146] Ibid., p. 81; Ekelhof 2018, p. 6
[147] NATO 2021b, p. 1-21.
[148] Similar to the 'Assessment' phase within the targeting cycle, many military frameworks emphasise the inclusion of civilian casualty tracking in after-action data collection; see, for example, US Department of Defense 2015, Sect. 5.11.1.3; NATO 2021b pp. 1-1 However, as noted above, the extent to which this practice is consistently implemented or prioritised remains uncertain.

been able to precisely identify that the misclassifications were triggered by a specific textile pattern viewed from a particular angle, as this would have required detailed technical analysis. However, even without such granular insights, component commanders could still have inferred from the pattern of misclassification reports that failures were linked to the local civilian population. Such a conclusion could likely have been reached before the one-month operation ended, allowing for a timely intervention. A horizontal communication mechanism is thus invaluable for enabling decentralised interventions at the operational level.

In addition to the above, vertical communication plays a critical role in facilitating more effective top-down interventions. In the *Drone* case, for example, a superior officer could have identified the pattern of errors based on reports provided by component commanders, enabling higher-level corrective measures. Vertical communication is also indispensable for the Iterative Review process. As noted earlier, Iterative Review is a *longitudinal* process: it relies on operational insights to determine whether prior legal assessments remain valid or require re-examination. However, *obtaining* such reliable and actionable data depends on effective coordination and cooperation from operational-level officers and soldiers.

Finally, AI presents novel opportunities to enhance and support Iterative Assessment by enabling more precise monitoring, data collection, and incident reconstruction.[149] For example, design standards could mandate that AI systems log all input-output pairings to facilitate post-action analysis,[150] similar to the black boxes used in commercial aircraft or the Aegis combat system.[151] Such data can then 'be analysed and allow[s] for improvement in decision sequences in the future'.[152] This approach effectively 'outsources' part of the Iterative Assessment workload to the AI system itself, integrating pre-installed monitoring and recording functionalities to enhance real-time oversight and post-deployment evaluation. By embedding these automated assessment mechanisms, the iterative framework could become more efficient, scalable, and appealing for military adoption.[153] Further research is needed to determine optimal design and integration strategies, as well as effective incentives to encourage militaries to voluntarily adopt such systems. Developing robust AI monitoring protocols could play a critical role in strengthening civilian protection by reducing the risk of avoidable 'accidents', which, under the right conditions, could ultimately be avoided in future AI-enabled operations.

5.2 Proactive Response

Having gained information that reasonably challenges the prior belief that an AI system's use was appropriate and lawful under the given circumstances, users are now obliged to take proactive action in the planning of future iterations—specifically, in subsequent operations or attacks involving the same AI system.[154]

As with Iterative Awareness, this requirement has both practical and legal foundations. From an operational perspective, for example, US doctrine recognises that '[a]rmy units can prevent civilian casualties by incorporating lessons learned from previous incidents, including near misses',[155] and that, in response, 'appropriate actions should be taken to reduce the risk of such incidents in the future'.[156] From a legal standpoint, this obligation is an integral part of a commander's precautionary duty to take appropriate measures to prevent foreseeable risks to civilians.[157] Proper implementation of Iterative Awareness (Sect. 5.1) acts as a necessary enabler: without it, inaction cannot be objectively judged as unreasonable, as the risk would not yet have been established as foreseeable for a reasonable commander in their position. This observation also has implications for any criminal liability one may want to ascribe to the decision maker

[149] Sassòli 2014, p. 326; Toscano 2015, p. 238.
[150] Williams 2017; see also European Committee on Crime Problems 2020.
[151] Aspin 1988.
[152] Meier 2019, p. 312.
[153] It has also been suggested that automated recording functions could mitigate the risk of subjectivity and collusion in investigations, as data is collected autonomously by the system: see Toscano 2015, p. 238.
[154] This principle applies across platforms as long as they operate on the same underlying algorithm. In the *Drone* case, for example, commanders would need to take action even if their own system had never malfunctioned, as the risks identified in their colleagues' drones would also apply to their own.
[155] Department of the Army Headquarters 2015 paras. 5-48.
[156] US Department of Defense 2015, Sect. 5.11.1.3. Note that 'reviews' as used in this quote does not refer to the general review obligation as discussed above, but rather the *ad hoc* assessments made after individual incidents.
[157] Margulies 2021, p. 177; Ekelhof 2018, p. 69; and API, above n. 1, art. 57(1). The US Law of War Manual categorises 'After-Action Assessments and Investigations' and 'Assessing the Risks to Civilians' under the header 'Feasible precautions in planning and conducting attacks': see US Department of Defense 2015, Sect. 5.11.1.3.

after the fact: if technically knowable risks remained factually unknown to the decision maker at the time of the decision, no *mens rea* can be established.[158]

The *appropriate* operational response to a newly identified risk is entirely contextual. Different commanders may adopt varying, yet equally valid, measures to address the perceived threat. In the *Drone* case:

> **Platoon Commander-A** may withhold the use of the drone entirely;
> **Platoon Commander-B** may order additional ISR collection and only deploy the drone if no local civilians are identified in the area; and
> **Platoon Commander-C** may opt to use the drone solely in a defensive posture (e.g., intercepting enemy counterattacks in situations where no local civilians are present to misidentify).

All three measures achieve the overarching goal of mitigating a known risk to the civilian population and would thus be considered appropriate responses based on the commanders' current knowledge of the situation.

By contrast, an *inappropriate* response would be, for example, for **Platoon Commander-D** to permit drone deployment based solely on visual confirmation that no local civilians are wearing distinct attire. This approach would be flawed given the information available at the time, as it rests on unverified assumptions. As previously noted, operational-level assessments will inevitably involve some degree of imprecision due to time and resource constraints. In the *Drone* case, the only verifiable common denominator, based on the misclassification reports, was the affected civilian demographic—any further specificity would have been speculative. While we, as external observers, know that the misclassification was caused by the clothing pattern, Platoon Commander-D would have no way of confirming this hypothesis without further evidence. Acting on such an unsubstantiated assumption would therefore be unreasonable. A reasonable commander would adopt the broadest plausible hypothesis—i.e., that the misclassification issue '*has something to do with the local civilian population*' in general—and implement precautionary measures accordingly pending further intelligence.[159]

In the *Targeting Adviser* case, the only *immediate* and appropriate response would likely have been to suspend the use of the DSS, as it is unlikely that the underlying pattern of failure (i.e., urgency in prompts influencing outputs) could have been identified so quickly. In the absence of further information, the only safe assumption would be that the DSS was not functioning reliably enough to justify continued deployment.

Now, consider a scenario in which a skilled technical team rapidly identifies the cause of failure: the DSS had developed an emergent ability to infer user emotions from chat prompts. What would be the reasonable course of action following this new information? One possible response would be to issue updated user instructions, such as '**Avoid abbreviations**' or '**Avoid these words: {quicky, faster, __}**'.[160]

This approach might allow the system to be reinstated under controlled conditions. However, a strong counterargument exists: the discovery of this one emergent behaviour suggests the potential existence of *other*, as yet unknown, emergent properties. If the system can infer urgency, what prevents it from also detecting and responding to emotions such as anger, enthusiasm, or desperation—potentially producing other undesirable or unpredictable outcomes?

Given this level of uncertainty, the most prudent course of action—aligned with the reasonable commander standard—would be to retire the system and recommend extensive testing to determine how different emotive cues influence its outputs. The core argument is that, in situations of uncertainty, a reasonable commander would err on the side of caution, implementing broad precautionary measures to mitigate a range of possible risks until further information is available to pinpoint the precise causes of failure.[161]

[158] See Kwik 2025b.
[159] IHL affords commanders 'reasonable latitude in the exercise of good faith judgment under the myriad circumstances and difficult conditions of combat'. However, this judgment must be based on a reasonable belief derived from the information available to them at the time: see Huffman 2012, p. 17; see also Haque 2012, p. 97.
[160] For more examples, see Kwik 2025a.
[161] This overall recommendation remains subject to context-specific considerations in the field and the judgment of a reasonable commander, particularly regarding the criticality of the capability to succeed in a given operation.

6 Visual Operational Guide

For ease of reference, Fig. 1 provides a visual operational guide to the key components of the Iterative Assessment framework, as discussed in Sects. 4 and 5.[162] It illustrates how each step is interconnected and how the process is continuously reinforced through new information gathered during deployment. This feedback loop enables the progressive refinement of both operational practices and legal assessment over time.

For any new system, **Iterative Review (Sect. 4)** provides an initial assessment of the system's lawfulness, which then informs its operational and tactical deployment. While the system is in use, **consistent and proactive reporting**—across both horizontal and vertical channels—enables Iterative Awareness (Sect. 5.1). This process can be further supported by automated recording and reporting functionalities embedded within the system itself. In response to emerging concerns or incident reports, operators are expected to take **Proactive Responses** (Sect. 5.2), based on feasibility and reasonableness.

Finally, the insights gathered during deployment feed back into the Iterative Review process, allowing for the continuous reassessment of whether the initial legal evaluation remains valid or requires revision.

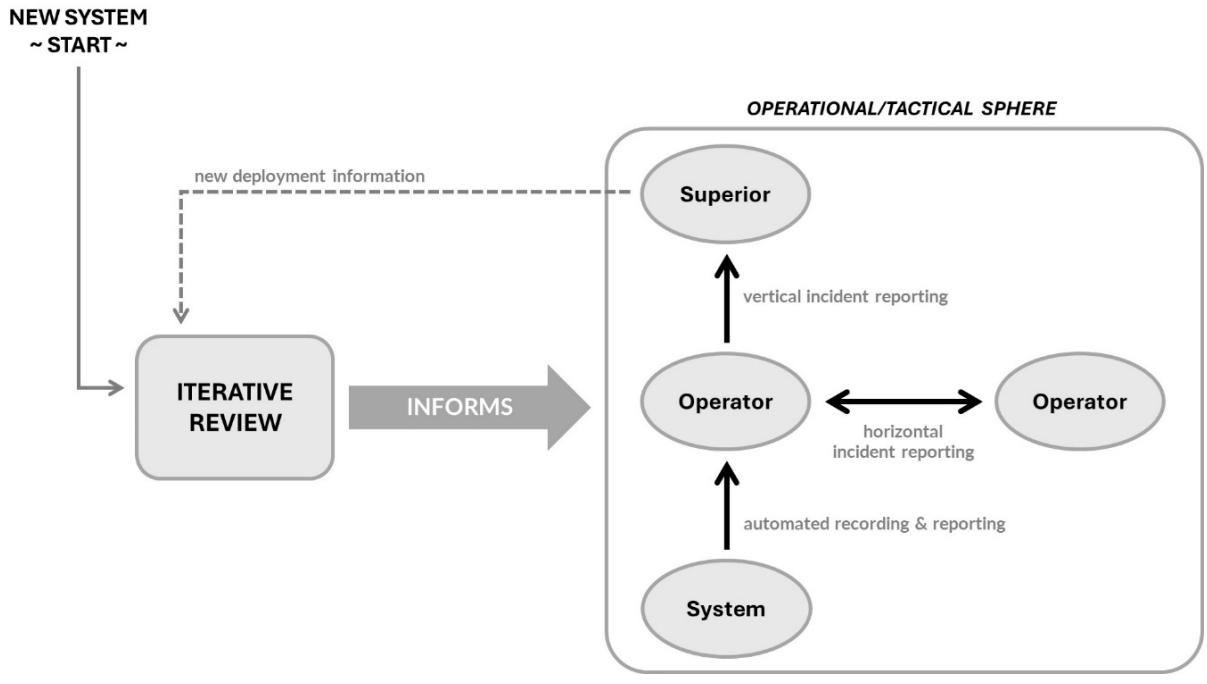

Figure 1: Visual Operational Guide to Iterative Assessment. *Source* The author.

[162] A more comprehensive visual summarising not only the Iterative Assessment steps, but also its underlying rationale as argued throughout this paper and comparisons with conventional assessment practices, can be viewed at https://jonathankwik.com/post/diagram-on-iterative-assessment. Accessed 20 June 2025.

7 Reflections and Recommendations

Warfare has always inherently been characterised by uncertainty. Incomplete knowledge and mistaken beliefs—even when justified at the time of decision making—can lead to harmful consequences on the battlefield. Recognising the realities of war, IHL judges' reasonableness based on the 'subjective perception of the decision maker at the time the decision was made'.[163] Under this standard, none of the actors involved in the *Drone* or *Targeting Adviser* cases—the commanders, operators or legal reviewers—acted unreasonably. Each relied on the information reasonably available to them at the time. However, as AI becomes increasingly integrated into military operations, such unforeseeable failures are likely to grow in frequency—even among good-faith belligerents. Pre-deployment controls alone cannot fully account for the genuinely unpredictable characteristics of military AI. While rigorous *ex ante* legal and technical reviews remain crucial, they cannot eliminate the possibility of unforeseen system failures in real-world conditions.

It is unrealistic to expect that initial accidents can be entirely avoided, and unfair to demand that AI users foresee the unforeseeable. In both the *Drone* and *Targeting Adviser* cases, no actor acted unreasonably from a legal perspective. Yet the outcomes were still suboptimal—had more robust safeguards been in place, further civilian harm could have been prevented:

> War inevitably involves death and destruction. The only way to avoid death and destruction in war is to avoid war. The fundamental purpose of [IHL] is to reduce net human suffering and net damage to civilian objects in armed conflict.[164]

IHL calls on belligerents to minimise *net* suffering—a principle that, in these cases, would mean intervening earlier to reduce the overall number of errors. Iterative Assessment provides a solution to meet this objective. It aligns with the spirit of the Rendulic Rule, avoiding retrospective blame while encouraging forward-looking improvements.[165] Instead of penalising decision makers for acting without perfect foresight, Iterative Assessment promotes the systematic collection of information to improve real-time decision making over the course of a system's usage.

As demonstrated in this paper, Iterative Awareness serves as a key enabler. Had horizontal reports from multiple platoons been available, reasonable commanders would likely have recognised the risks associated with drone operations in regions where distinct local attire was prevalent. Similarly, an effective civilian harm tracking mechanism would have alerted *Targeting Adviser* officers to the presence of an unanticipated variable affecting their DSS, undermining their previously justified confidence in its reliability. Once such critical information becomes available, the situation shifts: at that point, good-faith users would be expected to intervene to prevent further civilian harm. Consequently, it is justifiable to argue that any subsequent decision to continue using the AI system, despite these warnings, would be unreasonable.

Iterative Assessment is specifically designed to address 'known unknowns'—gaps in knowledge about system faults that can only be uncovered through real-world deployment. It does not seek to prevent initial failures but instead enables swift, decisive action to prevent their recurrence in future iterations. To this end, the

[163] Corn 2012, p. 451. See also Schmitt et al. 2017, p. 298.
[164] Fenrick 2005, p. 168.
[165] The Rendulic Rule establishes that 'commanders and personnel should be evaluated based on information reasonably available at the time of decision': see Kouba 2017, p. 10.

approach outlined in this paper represents best practice for both operational and post-action assessments and enhances civilian protection in *any* military operation, regardless of the capabilities involved.

Logistical and operational constraints may limit the comprehensive implementation of the framework. As emphasised from the outset, Iterative Assessment is an ideal rather than a legally binding rule. There are valid reasons why many of the recommended mechanisms—such as maintaining a permanent re-review structure post-deployment and integrating horizontal and vertical reporting systems—are rarely adopted in practice. These measures are time, resource, and labour intensive. For instance, conducting detailed civilian harm tracking after every attack may divert critical assets from other operational needs,[166] while establishing a robust reporting system requires dedicated communication networks and personnel. Nevertheless, militaries are encouraged to adopt Iterative Assessment to the extent feasible as it offers a structured means of mitigating the uncertainty that AI systems inevitably bring to the battlefield.

[166] Hathaway and Khan 2024, p. 55.

References

Abaimov S, Martellini M (2020) Artificial Intelligence in Autonomous Weapon Systems. In: Martellini M, Ralf T (eds) 21st Century Prometheus. Springer International Publishing, Cham, pp. 141-177

Abraham Y (2024) 'Lavender': The AI machine directing Israel's bombing spree in Gaza. https://www.972mag.com/lavender-ai-israeli-army-gaza. Accessed 15 April 2024

Adadi A, Berrada M (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE
DOI: 10.1109/ACCESS.2018.2870052

Aspin L (1988) Witness to Iran Flight 655. https://www.nytimes.com/1988/11/18/opinion/witness-to-iran-flight-655.html. Accessed 25 March 2025

Aspin L (1988) Witness to Iran Flight 655. New York Times, 18 November 1988

Australia (2006) Law of Armed Conflict. Defence Publishing Service, Canberra, Australian Defence Doctrine Publication 06.4. https://www.onlinelibrary.iihl.org/wp-content/uploads/2021/05/AUS-Manual-Law-of-Armed-Conflict.pdf. Accessed 9 May 2025

Australia (2018) The Australian Article 36 Review Process. CCW GGE on LAWS, UN Doc CCW/GGE.2/2018/WP.6. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS_August_Working%2Bpaper_Australia.pdf. Accessed 9 May 2025

Bakirtzis G, Carr S, Danks D, Topcu U (2022) Dynamic Certification for Autonomous Systems, arXiv:2203.10950v3 [cs.RO]
DOI: 10.48550/arXiv.2203.10950

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. Information Fusion
DOI: 10.1016/j.inffus.20112.012

Bartels R (2013) Dealing with the Principle of Proportionality in Armed Conflict in Retrospect: The Application of the Principle in International Criminal Trials. Israel Law Review
DOI: 10.1017/S0021223713000083

Belgian Armed Forces (2018) Procedure specifique / Specifieke procedure. DGJUR-SPS-CJBCEJ-CXX-001/LEGAD-Int, Ed 001/Rév 000, 21 Novembre 2018

Bommasani R et al. (2021) On the Opportunities and Risks of Foundation Models. arXiv:2108.07258v3 [cs.LG]
DOI: 10.48550/arXiv.21007258

Boothby WH (2016) Weapons and the Law of Armed Conflict, 2nd edn. Oxford University Press, Oxford

Boothby WH (2018) New Technologies and the Law in War and Peace. Cambridge University Press, Cambridge

Bothe M, Partsch KJ, Solf WA (eds) (2013) New Rules for Victims of Armed Conflict: Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949, 2nd edn. Martinus Nijhoff, Leiden

Boulanin V (2016) Mapping the Development of Autonomy in Weapon Systems: A Primer on Autonomy. https://www.sipri.org/sites/default/files/Mapping-development-autonomy-in-weapon-systems.pdf. Accessed 25 March 2025

Boulanin V, Goussac N, Bruun L, Richards L (2020) Responsible Military Use of Artificial Intelligence: Can the European Union Lead the Way in Developing Best

Practice. https://www.sipri.org/sites/default/files/2020-11/responsible_military_use_of_artificial_intelligence.pdf. Accessed 25 March 2025

Cambridge Dictionary (2025) Iteration. https://dictionary.cambridge.org/dictionary/english/iteration. Accessed 25 March 2025

Cherry J, Johnson D (2020) Maintaining Command and Control (C2) of Lethal Autonomous Weapon Systems: Legal and Policy Considerations. Southwestern Journal of International Law 27:1-27

Copeland DP (2014) Legal Review of New Technology Weapons. In: Nasu H, McLaughlin R (eds) New Technologies and the Law of Armed Conflict. T.M.C. Asser Press, The Hague, pp. 43-55

Corn GS (2012) Targeting, Command Judgement, and a Proposed Quantum of Information Component: A Fourth Amendment Lesson in Contextual Reasonableness. Brooklyn Law Review 77(2):437-498

Cranny-Evans S (2024) Sensor fusion: The future of land ISTAR?. European Security and Defence. https://euro-sd.com/2024/04/articles/37593/sensor-fusion-the-future-of-land-istar. Accessed 24 July 2024

Curtis E. Lemay Center (2019) Air Force Doctrine Publication 3-60 – Targeting. Curtis E Lemay Center. www.doctrine.af.mil/Doctrine-Publications/AFDP-3-60-Targeting. Accessed 5 July 2021

Daoust I, Coupland R, Ishoey R (2002) New wars, new weapons? The obligation of States to assess the legality of means and methods of warfare. International Review of the Red Cross 84:345-362

Das S (2021) Best Practices for Dealing with Concept Drift. Neptune.ai. https://neptune.ai/blog/concept-drift-best-practices. Accessed 30 October 2021

Defense Innovation Board (2019) AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Innovation Board. US Department of Defense. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. Accessed 2 August 2023

Deng B (2015) The Robot's Dilemma: Working out How to Build Ethical Robots Is One of the Thorniest Challenges in Artificial Intelligence. Nature 523:25-27

Department of the Army Headquarters (2015) Protection of Civilians. https://irp.fas.org/doddir/army/atp3-07-6.pdf. Accessed 25 March 2025

Dickson B (2022) AI scientists are studying the "emergent" abilities of large language models. https://bdtechtalks.com/2022/08/22/llm-emergent-abilities/. Accessed 27 January 2024

Ekelhof M (2018) Lifting the Fog of Targeting: "Autonomous Weapons" and Human Control through the Lens of Military Targeting. Naval War College Review 71(3):61-94

Ekelhof M, Paoli GP (2020) Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems. UNIDIR. https://unidir.org/publication/swarm-robotics-technical-and-operational-overview-of-the-next-generation-of-autonomous-systems. Accessed 27 January 2023

European Committee on Crime Problems (2020) Feasibility Study on a Future Council of Europe Instrument on Artificial Intelligence and Criminal Law. https://rm.coe.int/cdpc-2020-3-feasibility-study-of-a-future-instrument-on-ai-and-crimina/16809f9b60. Accessed 25 March 2025

Fenrick WJ (2005) International Humanitarian Law and Combat Casualties. European Journal of Population / Revue européenne de Démographie DOI: 10.1007/s10680-005-6421-y

Flournoy MA, Haines A, Chefitz G (2020) Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, Including Deep Learning Systems. https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf. Accessed 25 March 2025

Fromm J (2005) Types and Forms of Emergence. arXiv:nlin/0506028 [nlin.AO]
DOI: 10.48550/arXiv.nlin/0506028

Fry J C (2006) Contextualized Legal Reviews for the Methods and Means of Warfare: Cave Combat and International Humanitarian Law. Columbia Journal of Transnational Law 44:453-519

Ganguli D et al. (2022) Predictability and Surprise in Large Generative Models. arXiv:2202.07785v2 [cs.CY]
DOI: 10.48550/arXiv.2202.07785

Geirhos R, Medina Temme CR, Rauber J, Schütt HH, Bethge M, Wichmann FA (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231v3 [cs.CV]
DOI: 10.48550/arXiv.1808.08750

Gilmer J, Adams RP, Goodfellow I, Andersen D, Dahl GE (2018) Motivating the Rules of the Game for Adversarial Example Research. arXiv:1807.06732 [cs.LG]
DOI: 10.48550/arXiv.1807.06732

Goussac N (2019) Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-Fighting. ICRC Humanitarian Law & Policy. https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting. Accessed 26 May 2021

Government Accountability Office (2023) Science & Tech Spotlight: Drone Swarm Technologies. https://www.gao.gov/assets/gao-23-106930.pdf. Accessed 25 April 2024

Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS) (2021) Draft Report of the 2021 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. CCW/GGE1/2021/CRP1

Guinness H (2024) What Are AI Hallucinations—and How Do You Prevent Them? https://zapier.com/blog/ai-hallucinations/. Accessed 24 March 2025

Gunnell M (2023) Emergent Behavior in AI. https://www.techopedia.com/definition/emergent-behavior. Accessed 27 January 2024

Haque AA (2012) Killing in the Fog of War. Southern California Law Review 86(1):63-116

Hathaway OA, Khan A (2024) 'Mistakes' in War. University of Pennsylvania Law Review 173:1-88

Henderson I (2009) The Contemporary Law of Targeting: Military Objectives, Proportionality and Precautions in Attack under Additional Protocol I. Martinus Nijhoff, Leiden

Hendrycks D, Liu X, Wallace E, Dziedzic A, Krishnan R, Song D (2020) Pretrained Transformers Improve Out-of-Distribution Robustness. arXiv:2004.06100 [cs.CL]
DOI: 10.48550/arXiv.2004.06100

Herbach J (2012) Into the Caves of Steel: Precaution, Cognition and Robotic Weapon Systems Under the International Law of Armed Conflict. Amsterdam Law Forum 4(3):3-20

Holland Michel A (2020) The Black Box, Unlocked: Predictability and Understandability in Military AI. UNIDIR, Geneva. https://unidir.org/files/2020-09/BlackBoxUnlocked.pdf. Accessed 2 August 2023

Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659) (2020) Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility. Publication Office of the European Union, Luxembourg

Huffman WB (2012) Margin of Error: Potential Pitfalls of the Ruling in The Prosecutor v. Ante Gotovina. Military Law Review 211:1-56

ICRC (2006) A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977. https://international-

review.icrc.org/sites/default/files/irrc_864_11.pdf. Accessed 25 March 2025

ICRC (2016) Views of the International Committee of the Red Cross (ICRC) on autonomous weapon system. https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system. Accessed 25 March 2025

ICRC (2018) 'Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?'. Convention on Certain Conventional Weapons (CCW), Group of Governmental Experts on Lethal Autonomous Weapons Systems (Geneva, 3 April 2018) https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control. Accessed 20 June 2025

ICRC (2019) 'Statement of the International Committee of the Red Cross (ICRC) under Agenda Item 5(B)'. Convention on Certain Conventional Weapons (CCW), Group of Governmental Experts on Lethal Autonomous Weapons Systems (Geneva, 25-29 March 2019) https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/CCW%2BGGE%2BLAWS%2BICRC%2Bstatement%2Bagenda%2Bitem%2B5b%2B26%2B03%2B201pdf. Accessed 9 May 2025

ICRC and Geneva Academy (2024) Expert Consultation Report on AI and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts. www.geneva-academy.ch/joomlatools-files/docman-files/Artificial%20Intelligence%20And%20Related%20Technologies%20In%20Military%20Decision-Making.pdf. Accessed 9 January 2025

Jenks C, Liivoja R (2018) Machine Autonomy and the Constant Care Obligation. ICRC Humanitarian Law & Policy. https://blogs.icrc.org/law-and-policy/2018/12/11/machine-autonomy-constant-care-obligation/. Accessed 28 May 2021

Jensen ET (2021) Autonomy and Precautions in the Law of Armed Conflict. In: Liivoja R, Väljataga A (eds) Autonomous Cyber Capabilities Under International Law. NATO CCDCOE Publications, Tallinn, pp. 181-205

Jevglevskaja N (2018) Weapons Review Obligation Under Customary International Law. International Law Studies 94:186-221

Kalmanovitz P (2016) Judgment, Liability and the Risks of Riskless Warfare. In: Bhuta N et al. (eds) Autonomous Weapons Systems. Cambridge University Press, Cambridge, pp. 145-163

King TC, Aggarwal N, Taddeo M, Floridi L (2020) Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. Science and Engineering Ethics DOI: 10.1007/s11948-018-00081-0

Klonowska K (2022) Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare. In: Gill T D, Geiß R, Krieger H, Mignot-Mahdavi R (eds) Yearbook of International Humanitarian Law, Volume 23 T.M.C. Asser Press, The Hague, pp. 123-156

Kouba MD (ed) (2017) Operational Law Handbook. 17th ed. International and Operational Law Department the Judge Advocate General's Legal Center and School, Charlottesville

Kwik J (2022) A Practicable Operationalisation of Meaningful Human Control. Laws DOI: 10.3390/laws11030043.

Kwik J (2024a) Lawfully Using Autonomous Weapon Technologies. TMC Asser Press, The Hague DOI: 10.1007/978-94-6265-631-4

Kwik J (2024b) Is wearing these sunglasses an attack? Obligations under IHL related to anti-AI countermeasures. International Review of the Red Cross. 106(926):732-75 DOI: 10.1017/S1816383124000067

Kwik J (2025a) Digital yes-men: How to deal with sycophantic military AI?. Global Policy (forthcoming 2025)

Kwik J (2025b) 'I Plead Ignorance': Autonomous Weapons and Criminal Liability for Not Knowing the

Knowable. International Law Studies (forthcoming 2025)

Kwik J, Van Engers T (2021) Algorithmic Fog of War: When Lack of Transparency Violates the Law of Armed Conflict. Journal of Future Robot Life DOI: 10.3233/FRL-200019

Kwik J, Zwanenburg M, Took C and Aarts J (2025) Controlling Military Artificial Intelligence: Harnessing Rules of Engagement and Military Directives. Asser Policy Brief 2025–01. https://papers.ssrn.com/abstract=5132731. Accessed 12 February 2025

Leslie D (2019) Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector. https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf. Accessed 25 March 2025

Liu H-Y (2016) Refining Responsibility: Differentiating Two Types of Responsibility Issues Raised by Autonomous Weapons Systems. In: Bhuta N, Beck N, Geiß R, Liu H-Y, Kreß C (eds) Autonomous Weapons Systems. Cambridge University Press, Cambridge, pp. 325-344

Lohn AJ (2020) Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance. arXiv:20000802 [cs.LG] DOI: 10.48550/arXiv.20000802

Margulies P (2021) A Moment in Time: Autonomous Cyber Capabilities, Proportionality and Precautions. In: Liivoja R, Väljataga A (eds) Autonomous Cyber Capabilities under International Law. NATO CCDCOE Publications, Tallinn, pp. 152-180

McClelland J (2003) The Review of Weapons in Accordance with Article 36 of Additional Protocol I. International Review of the Red Cross 850:397-415

McGrayne SB (2011) The Theory That Would Not Die. Yale University Press, New Haven

Meier MW (2019) Lethal Autonomous Weapons Systems: Is It the End of the World as We Know It . . . Or Will We Be Just Fine?. In: Williams W S and Ford C M (eds) Complex Battlespaces: The Law of Armed Conflict and the Dynamics of Modern Warfare. Oxford University Press, Oxford, pp. 289-316

Meng F-J, Li Y-Q, Shao F-M, Yuan G-H, Dai J-Y (2022) Visual-simulation region proposal and generative adversarial network based ground military target recognition. Defence Technology DOI: 10.1016/j.dt.2021.07.001

Mikhailov DI (2023) Optimizing National Security Strategies through LLM-Driven Artificial Intelligence Integration. arXiv:2305.13927 [cs.CY] DOI: doi.org/10.48550/arXiv.2305.13927

Mogul JC (2005) Emergent (Mis)behavior vs. Complex Software Systems. HP Invent, HPL-2006-2. https://www.hpl.hp.com/techreports/2006/HPL-2006-2.pdf. Accessed 24 January 2024

Molnar C (2019) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lean Publishing

Moyes R (2019) Target Profiles: An Initial Consideration of "Target Profiles" as a Basis for Rule-Making in the Context of Discussions on Autonomy in Weapons Systems. https://article36.org/wp-content/uploads/2019/08/Target-profiles.pdf. Accessed 25 March 2025

NATO (2019a) Allied Joint Doctrine for the Planning of Operations. NATO Standard, AJP-5, Edition A Version 2, May 2019

NATO (2019b) Allied Joint Doctrine for the Conduct of Operations. NATO Standard, AJP-3, Edition C Version 1, February 2019

NATO (2021a) Summary of the NATO Artificial Intelligence Strategy. https://www.nato.int/cps/en/natohq/official_texts_187617.htm. Accessed 6 September 2023

NATO (2021b) Allied Joint Doctrine for Joint Targeting. Edition B Version 1 (November 2021) AJP-39, NATO Standardization Office (NSO)

Navarro I, Matía F (2013) An Introduction to Swarm Robotics. Robotics DOI: 10.5402/2013/608164

Nelson K, Corbin G, Blowers M (2014) Evaluating Data Distribution and Drift Vulnerabilities of Machine Learning Algorithms in Secure and Adversarial Environments. In: Blowers M, Williams J (eds) Proceedings Volume 9119, Machine Intelligence and Bio-inspired Computation: Theory and Applications VIII. P 911904
DOI: 10.1117/12.2053045

Ngo R, Chan L, Mindermann S (2022) The Alignment Problem from a Deep Learning Perspective. arXiv:22000626v6 [cs.AI]
DOI: 10.48550/arXiv.22000626

Oeter S (2020) Specifying the Proportionality Test and the Standard of Due Precaution: Problems of Prognostic Assessment in Determining the Meaning of "May Be Expected" and "Anticipated". In: Kreß C, Lawless R (eds) Necessity and Proportionality in International Peace and Security Law. Oxford University Press, Oxford, pp. 343-366

Oniani D, Hilsman J, Peng Y, Poropatich R K, Pamplin J C, Legault G L, Wang Y (2023) From Military to Healthcare: Adopting and Expanding Ethical Principles for Generative Artificial Intelligence. arXiv:2308.02448 [cs.CY]
DOI: doi.org/10.48550/arXiv.2308.02448

Parks WH (2005) Conventional Weapons and Weapons Reviews. Yearbook of International Humanitarian Law 8:55-142

Perez E et al. (2022) Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251 [cs.CL]
DOI: 10.48550/arXiv.2212.09251

Puscas I (2023) AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures. https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/. Accessed 9 May 2025

Rosén F (2014) Extremely Stealthy and Incredibly Close: Drones, Control and Legal Responsibility. Journal of Conflict and Security Law
DOI: 10.1093/jcsl/krt024

Rosenfeld A (2021) Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th international conference on autonomous agents and multiagent systems, pp. 45-50. https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p45.pdf. (240803) Accessed 3 August 2024

Roy A (2018) National Strategy for Artificial Intelligence #AIFORALL. National Institution for Transforming India Aayog, New Delhi. https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf. Accessed 9 May 2025

Russell SJ, Norvig P (2021) Artificial Intelligence: A Modern Approach, 4th edn. Pearson, New Jersey

Russell S, Dewey D, Tegmark M (2015) Research Priorities for Robust and Beneficial Artificial Intelligence. Future of Life Institute. https://futureoflife.org/open-letter/ai-open-letter/. Accessed 2 February 2021

Safi M (2019) Are Drone Swarms the Future of Aerial Warfare? www.theguardian.com/news/2019/dec/04/are-drone-swarms-the-future-of-aerial-warfare. Accessed 2 August 2023

Sandoz Y, Swinarski C, Zimmerman B (1987) Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 194 Martinus Nijhoff, Geneva

Sassòli M (2014) Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified. International Law Studies 90:308-340

Sayler KM (2020) Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems. Congressional Report Service

Schaeffer R, Miranda B, Koyejo S (2023) Are Emergent Abilities of Large Language Models a Mirage?. arXiv:2304.15004v2 [cs.AI]
DOI: doi.org/10.48550/arXiv.2304.15004

Scharre PD (2016) Centaur Warfighting: The False Choice of Humans vs. Automation. Temple International & Comparative Law Journal 30(1):151-165

Schmitt MN (2010) Targeting in Operational Law. In: Gill T D, Fleck D (eds) The Handbook of the International Law of Military Operations. Oxford University Press, Oxford, pp. 245-275

Schmitt MN, Biller J, Fahey SC, Goddard DS, Highfill C, Ohlin JD, May L, Finkelstein C (2017) Joint and Combined Targeting: Structure and Process. In: Ohlin J D, May L, Finkelstein C (eds) Weighing Lives in War, vol 1. Oxford University Press, Oxford, pp. 298-324. DOI: 10.1093/oso/9780198796176.003.0014

Schmitt MN, Schauss M (2019) Uncertainty in the Law of Targeting: Towards a Cognitive Framework. Harvard National Security Journal 10:148-194

Schuller AL (2017) At the Crossroads of Control: The Intersection of Artificial Intelligence in Autonomous Weapon Systems with International Humanitarian Law. Harvard National Security Journal 8:379

Sharma M et al. (2023) Towards Understanding Sycophancy in Language Models. arXiv:2310.13548v3 [cs.CL]
DOI: doi.org/10.48550/arXiv.2310.13548

Gilmer J, Metz L, Faghri F, Schoenholz S S, Raghu M, Wattenberg M, Goodfellow I (2018) Adversarial Spheres. arXiv:1801.02774 [cs.CV]
DOI: 10.48550/arXiv.1801.02774

Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M (2023) Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. International Journal of Human–Computer Interaction
DOI: doi.org/10.1080/10447318.2022.2101698

Steinhardt J (2022) Future ML Systems Will Be Qualitatively Different. Bounded Regret, https://bounded-regret.ghost.io/future-ml-systems-will-be-qualitatively-different. Accessed 27 January 2024

Stewart DM (2011) New Technology and the Law of Armed Conflict. International Law Studies 87:271-298

Sullivan S, Ricket I (2024) Targeting in the Black Box. In: Kwan C, Lindström L, Giovannelli D, Podiņš K, Štruc D (eds) CyCon 2024: Over the Horizon. Tallinn, NATO CCD COE, pp. 207-220. https://ccdcoe.org/uploads/2024/05/CyCon_2024_book.pdf#page=215. Accessed 25 March 2025

Switzerland (2016) Towards a "compliance-based" approach to LAWS. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWS%2BMX_Countrypaper%2BSwitzerland.pdf. Accessed 25 March 2025

Thorne JG (2020) Warriors and War Algorithms: Leveraging Artificial Intelligence to Enable Ethical Targeting. Technical Report, 14-05-2020. https://apps.dtic.mil/sti/citations/AD1104171. Accessed 3 July 2021

Tolk A (2015) Merging Two Worlds: Agent-Based Simulation Methods for Autonomous Systems. In: Williams A P, Scharre P D (eds) Autonomous Systems: Issues for Defence Policymakers. NATO, The Hague, pp. 291-317

Toscano C P (2015) "Friend of Humans": An Argument for Developing Autonomous Weapons Systems. Journal of National Security Law & Policy 8(1):189-246

Trusilo D (2023) Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability. Journal of Military Ethics 22:2–17 https://www.tandfonline.com/doi/full/10.1080/15027570.2023.2213985

UK Ministry of Defence (2016) UK Weapon Reviews. https://www.gov.uk/government/publications/uk-weapon-reviews. Accessed 3 August 2024

Undersecretary of Defense for Acquisition, Technology, and Logistics (2012) The Role of Autonomy in DoD Systems. https://irp.fas.org/agency/dod/dsb/autonomy.pdf. Accessed 25 March 2025

US Air Force Office of the Chief Scientist (2015) Autonomous Horizons: System Autonomy in the Air Force—A Path to the Future, Volume I: Human-Autonomy Teaming. https://www.af.mil/Portals/1/documents/SECAF/AutonomousHorizons.pdf. Accessed 25 March 2025

US Department of Defense (2015) Law of War Manual, Updated July 2023. US Department of Defense, Washington D.C.

US Department of Defense (2023) Autonomy in Weapon Systems. US DoD Directive 3000.09

US Joint Chiefs of Staff (2018) Counterinsurgency. Joint Publication 3-24, 25 April 2018, Validated 30 April 2021

Van den Boogaard JC, Roorda MP (2021) "Autonomous" Weapons and Human Control. In: Bartels R (eds) Military Operations and the Notion of Control Under International Law. TMC Asser Press, The Hague, pp. 421-437

Van der Waa J, Nieuwburg E, Cremers A, Neerincx M (2021) Evaluating XAI: A Comparison of Rule-Based and Example-Based Explanations. Artificial Intelligence 291:103404. https://www.sciencedirect.com/science/article/pii/S0004370220301533

Wallace D (2018) Cyber Weapon Reviews under International Humanitarian Law: A Critical Analysis. Tallinn Paper no 11, https://ccdcoe.org/uploads/2018/10/TP-11_2018.pdf. Accessed 3 August 2024

Williams R (2017) Lords select committee, artificial intelligence committee, written evidence. AIC0206. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70496.html#_ftn13. Accessed 7 May 2022

Wilson C (2020) Artificial Intelligence and Warfare. In: Martellini M, Trapp R (eds) 21st Century Prometheus. Springer International Publishing, Cham, pp 125-140 DOI: 10.1007/978-3-030-28285-1_7

Yadron D, Tynan D (2016) Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode. www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk. Accessed 25 March 2025

About the author

Dr Jonathan Kwik

Related works at:
https://jonathankwik.com/publications/

About this project

ELSA Lab Defence

ELSA Lab Defence is developing a future-proof, independent, and consultative ecosystem for the responsible use of AI within the defence domain. AI plays a vital role in defence, enhancing military efficiency and addressing challenges like adversarial AI. However, its integration raises ethical, legal, and societal concerns, including maintaining human control, dignity, and legal boundaries. ELSA Lab Defence aims to develop a responsible framework for AI in defence by focusing on context-dependent analysis, ethical design, and human-machine collaboration. It explores the perception of military AI among society and defence personnel, monitoring global trends to ensure the responsible, value-driven use of AI technologies in defence applications.

Link to Project: https://elsalabdefence.nl/

ELSA LAB DEFENCE

About the Asser Institute, Centre for International and European law

The Asser Institute's mission is to contribute to the development of international and European public and private law. We achieve this by:

- **Independent legal research:** We conduct fundamental, policy-oriented, and applied legal research in international and European public and private law.
- **Knowledge dissemination:** We initiate and facilitate academic and expert meetings, (professional) education, and public events aimed at disseminating knowledge of international and European public and private law. We further share our legal knowledge by adding to the public debate.

ASSER INSTITUTE
Centre for International & European Law

T.M.C. Asser Instituut
R.J. Schimmelpennincklaan 20-22
2517 JN The Hague
The Netherlands

P.O. Box 30461
2500 GL The Hague
The Netherlands

+31 (0)70 342 03 00
info@asser.nl

asser.nl