

SAFETY, LEGAL, AND ETHICAL IMPLICATIONS OF ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN

EDITED BY: ZENA ASSAAD, JESSICA DORSEY
AND NEIL RENIC

AUSTRALIAN ARMY RESEARCH CENTRE

/ Australian Army Collected Papers Series





SAFETY, LEGAL, AND ETHICAL IMPLICATIONS OF ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN

EDITED BY: ZENA ASSAAD, JESSICA DORSEY
AND NEIL RENIC

AUSTRALIAN ARMY RESEARCH CENTRE
/ Australian Army Collected Papers Series

© Commonwealth of Australia 2026

This publication is copyright. Apart from any fair dealing for the purpose of study, research, criticism or review (as permitted under the *Copyright Act 1968*), and with standard source credit included, no part may be reproduced by any process without written permission.

The views expressed in this publication are those of the author(s) and do not necessarily reflect the official policy or position of the Australian Army, the Department of Defence or the Australian Government.

ISSN (Print) 1448-2843

ISSN (Digital) 2200-0992

DOI: 10.61451/2675150

All enquiries regarding this publication should be forwarded to the Director of the Australian Army Research Centre.

To learn about the work of the Australian Army Research Centre visit

<https://researchcentre.army.gov.au>

Cover image: Australian Army soldiers from 8th/7th Battalion, Royal Victoria Regiment, and Royal Australian Air Force aviators from 1st Security Forces Squadron took part in an enhanced combat shooting course at Puckapunyal, Victoria, in June 2024.

CONTENTS

Safety, Legal and Ethical Implications of Artificial Intelligence in the Military Domain: An Australian Army Perspective	1
<i>Zena Assaad, Jessica Dorsey and Neil Renic</i>	
AI-DSS and the Erosion of Technical Integrity: Rethinking Human Decision-Making in Modern Warfare.....	4
<i>Zena Assaad and Adam Hepworth</i>	
Pathways Towards Ensuring Human Control and Judgement Across the Military AI Lifecycle: An Australian Army Perspective	28
<i>Ingvild Bode and Anna Nadibaidze</i>	
Responsible Procurement of Military Artificial Intelligence in Australia: A Window of Opportunity	50
<i>Netta Goussac and Vincent Boulanin</i>	
The Role of Article 36 Legal Reviews Across the Military AI Assurance Lifecycle	66
<i>Damian Copeland</i>	
TEVV and the Responsible and Lawful Use of Military AI Systems.....	92
<i>Manoj Harjani</i>	
Digitising the Commander's Intent Towards Effective Governance of Decentralised Human–AI Teams	119
<i>Adam Hepworth and Hussein Abbass</i>	
Governing Joint Innovation: Legal Review and Emerging Technologies under AUKUS Pillar 2	138
<i>Natalia Jevglevskaja and Renato Wolf</i>	
Predictability in the Governance of Autonomous Weapon Systems	166
<i>Rain Liivoja</i>	

Architecting Trust: A ‘Trust by Design Plus’ Framework for Military AI189
Branka Marijan

Military AI Systems and the Design of Lifecycle-Sensitive
Governance Frameworks 207
Rob McLaughlin and Conor McLaughlin

Built In or Bolted On: Industry’s Role in Designing
Human–Machine Interaction in AI-Enabled Land Warfare 225
Lena Trabucco

Industry Development of Military AI: Operationalising International
Humanitarian Law Across the Lifecycle 245
Taylor Kate Woodcock

/ SAFETY, LEGAL AND ETHICAL IMPLICATIONS OF ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN: AN AUSTRALIAN ARMY PERSPECTIVE

Zena Assaad, Jessica Dorsey and Neil Renic

The integration of artificial intelligence into a range of military systems is no longer the subject of theoretical speculation. States are, at this moment, designing, testing, procuring and deploying AI-enabled and autonomous capabilities across all battlefield domains. These technological and military advances are taking place within an insecure geopolitical climate—intensified great power contestation; a resurgence of civilisational politics; destabilising arms racing; and the return of large-scale, inter-state war. Amidst such conditions, calls are growing for middle powers like Australia to evaluate their approach to military AI—accelerate innovation, compress regulatory processes, and optimise warfighting to better position itself against the security challenges facing the region.

This narrative leaves a lot unaddressed, including a range of questions with genuine moral weight. Firstly, is the military AI arms race as critical and seemingly inevitable as frequently described? If it is, then what does this actually license Australian policymakers to do in this space? Are existing technical safeguards sufficient or in need of reform? What are the unintended but foreseeable negative effects of speeding and scaling up our technological acquisition? Within the increasingly systematic, data-infused, sensor-saturated battlefield we are building towards, can we realistically maintain human operators as something more than morally inert actioning devices? These questions do not fall away, regardless of how seemingly dire the political and military stakes may be.

Just as importantly, who ought to bear the costs for the choices we are making; for our greater appetite for innovation, experimentation and risk; for the uncertainties we are accepting, and in some cases, magnifying? Too often in war, it is the innocent. Responsible governance of this technology requires meaningful reflection about the innocent civilians most likely and least able to bear the costs when our military-technological policies go awry.

There are also considerations around how existing legal frameworks can be applied to technologies that are changing the circumstances in which military decisions are made. How legal requirements are to be interpreted and implemented as decision-making becomes increasingly dependent on systems that can process vast quantities of data, generate recommendations, and in some circumstances act with limited (or no) human intervention is becoming more critical by the day. These developments call into question the meaning and practical operation of human judgement and control, and make us critically inquire how legal responsibility can be maintained when the processes through which decisions are reached become increasingly complex, distributed, and technologically mediated.

For Australia, these problems are further shaped by the increasingly collaborative nature of defence technology development, including through AUKUS, a factor addressed in these collected papers. The development and integration of AI-enabled capabilities with allies and partners presents significant opportunities, while also creating challenges concerning interoperability, differing legal and regulatory requirements, assurance standards, and the allocation of responsibility across national and organisational boundaries. More fundamentally, AI integration may alter the 'choice architecture' within which military decisions are made, in so doing shaping which options are identified, how information is presented, the speed at which decisions must be taken, and the opportunities available for appropriate levels of human intervention and compliance with international law. The legal implications therefore extend beyond the question of whether particular uses of AI are lawful. These systems also introduce socio-technical implications concerning whether the systems, processes and institutional arrangements through which military decisions are made are capable of sustaining the exercise of lawful judgement and preserving meaningful accountability at various levels as the character of military decision-making changes.

The contributions in this compendium take seriously these questions, detailing how the Australian military can build meaningful assurance into AI-enabled systems across the entire lifecycle: research and design, procurement, testing, legal review, deployment, and use. These contributions are not agnostic as to the hard choices and difficult trade-offs facing Australia within this security environment. But nor do they accept that the same pressures justify an abandonment of safety, legal, moral and political considerations.

We would like to acknowledge the time and effort of the following people who reviewed and provided feedback on the papers within this series: Mel McDowell, Hannah Crothers, Nicholas Carroll, Laura Bound, Simon McKenzie, Austin Wyatt, Chris Kourloufas, Lauren Sanders, Emily Clarke and Callum Hamilton.

About the Authors

Dr Zena Assaad is an Associate Professor in Engineering at the Australian National University (ANU) and the Deputy Director of the ANU Defence Institute. She is currently a fellow with the UN Institute for Disarmament Research (UNIDIR) leading research on the decommissioning of military AI systems. She has held fellowships with the Australian Army Research Centre and Trusted Autonomous Systems. Her research explores the safety of critical and complex technologies in safety-critical domains, with a focus on systems engineering practices. She is currently focused on topics related to autonomy and AI in military systems. Dr Assaad is a former member of the expert advisory group for the global commission on responsible AI in the military domain (GC REAIM). She is currently an Ambassador for the Responsible by Design initiative of Article 36 Legal. She also serves on the independent advisory board on Legal Reviews as a technical expert. She is also the host of the Responsible Bytes podcast.

Dr Jessica Dorsey is an Assistant Professor of International Law and Co-Director of the Realities of Algorithmic Warfare Research Platform at Utrecht University. She is also a member of the expert advisory group of the Global Commission on Responsible AI in the Military Domain (GC REAIM) and an Ambassador to the Responsible by Design initiative of Article 36 Legal, which aims to strengthen national legal review processes and promote IHL compliance throughout the lifecycle of AI-based military technologies. Her current research focuses on the legitimacy of military targeting operations in light of increasing autonomy in warfare, with a focus on the protection of civilians in armed conflict. Jessica is also a renowned legal scholar and practitioner on issues related to the use of force, especially in the context of drone warfare, having worked in the field for more than 15 years. She is an executive board member of Airwars and the Managing Editor of the international law weblog *Opinio Juris*.

Dr Neil Renic is a lecturer in ethics at the University of New South Wales, a fellow at the Centre for Military Studies at the University of Copenhagen, and member of the International Committee for Robot Arms Control (ICRAC). He is also the Associate Director of the Military Ethics Research Lab and Innovation Network (MERLIN). He specialises in the changing character of war, the ethics of killing, and emerging military technologies. He is the author of “Asymmetric Killing: Risk Avoidance, Just War, and the Warrior Ethos” (Oxford University Press 2020).

/ AI-DSS AND THE EROSION OF TECHNICAL INTEGRITY: RETHINKING HUMAN DECISION- MAKING IN MODERN WARFARE

Zena Assaad and Adam Hepworth

‘Someone decided to compress the kill chain... Calling it an “AI problem” gives those decisions, and those people, a place to hide.’

Kevin T Baker, *The Guardian*, 26 March 2026

Introduction

The adage of military artificial intelligence (AI) no longer being an imagined reality has become a standard opening remark for much of the discourse on the topic of AI in military operations. Current conflicts in the Ukraine, Palestine and Iran serve as a backdrop for this, demonstrating multiple accounts of AI use in active conflicts. This has shifted the discourse away from hypothetical instantiations to real-world case studies, with aspects of these case studies involving the use of AI decision-support systems (AI-DSS) for targeting operations.

AI-DSS are currently being used in the Israel–Gaza war¹ and the Russia–Ukraine war² and most recently in support of Operation Epic Fury between the United States, Israel and Iran.^{3 4} AI-DSS are systems designed to support decision-making by collating and analysing large datasets, shortening decision loops⁵ and generating decision-support outputs.⁶ These tools have a wide range of uses, which can be categorised into three broad functions:

- description, synthesis and analysis, which involves collecting, organising and presenting data
- prediction and extrapolation, which entails identifying patterns and trends in data, including forecasting possible outcomes and their probabilities
- recommendations, which focus on recommending the most effective course(s) of action.⁷

Specific operational examples of AI-DSS used for targeting include Lavender (target prioritisation), The Gospel (intelligence fusion) and Where's Daddy (location tracking), which have been used by the Israel Defence Forces (IDF) in Gaza.⁸ Lavender uses machine learning to assign a numerical score indicating the likelihood of a person being a member of an armed group. The Gospel processes surveillance data to generate and categorise targets. Where's Daddy utilises location data from connected devices to detect and locate people identified as military targets.

More recently, there have been reports indicating the use of Palantir's Maven Smart System, a machine-learning and sensor-fusion command and control platform, as well as Anthropic's Claude, a frontier large language model, in a decision-support capacity by the US military for targeting operations in Iran as part of Operation Epic Fury.⁹ Public reports suggest Claude was being used for 'intelligence assessments, including for target identification and simulating battle scenarios' in the lead-up to strikes on Iran.

The IDF claims that the use of AI-DSS has increased operational precision and accuracy,^{10 11} although reports of increased civilian death tolls in Gaza provide contradictory claims.¹² This outcome is, in part, due to the limited accuracy, validation and reliability of these systems in real-world scenarios, as well as the level of procedural integration and oversight of targeting decisions.¹³ This highlights the impact of upstream decision-making in capability development intersecting with operational practices. Technology capabilities are not deployed in a vacuum; rather they are implemented into existing military systems and operational procedures, some of which they complement and others with which they conflict. These points of conflict are often identified and mitigated in both acquisition and operational processes; however, as the speed of fielding is prioritised, established processes can be narrowed or even overlooked.

In this paper we argue that the narrowing and omission of critical military processes results in an accumulative degradation of human decision-making over the lifecycle of an AI-enabled system, leading to eroded technical integrity and compromised operational effectiveness. We frame operational effectiveness in this paper through the lens of verification leading to protection, including the safety of non-combatants and combatants, consistent with the international humanitarian law (IHL) principles of distinction (distinguishing between combatants and non-combatants) and proportionality (ensuring disproportionate impacts to civilians during military operations are minimised). Through this lens, an AI-DSS may be functioning 'as designed' from an operational verification perspective, while still producing outcomes incompatible with objective IHL distinction and proportionality assessments.

We posit that human decision-making upstream of the point of use—in capability development through requirements setting, acquisition, design, test and evaluation, regulatory and standards compliance including Article 36 legal reviews, and the

assignment of accountable officers—is critical to technical integrity and operational effectiveness for AI-DSS when used in operational targeting processes.

This position aligns with Australian responsible AI frameworks as presented in Australia's Policy Settings for Responsible Use of Artificial Intelligence in Defence (2025), which establishes a values-based approach through three obligations for AI use: lawfulness, adherence to values-based principles, and proportionate controls.¹⁴ The policy requires Defence personnel to consider legal obligations at all stages of decision-making to acquire and use AI. While this is specific to legal obligations, it should be noted that the requirements stage establishes where principles of responsible behaviour must first be embedded, including technical requirements of which some fall under legally mandated regulations. Goussac and Boulanin¹⁵ articulate a similar approach, emphasising a need to consider principles of responsible behaviour throughout the acquisition lifecycle, rather than attempting to 'retrofit' them at a later stage.

While this paper focuses on human decision-making, comparative literature on human agency should also be noted. Conn & Bode suggest that upstream agency could be eroded for AI-enabled systems, noting:

[M]ost of these ethical and legal discussions have revolved primarily around the point of use of a hypothetical AIS,¹⁶ and in doing so, one critical component still remains under-appreciated: human decision-making across the full timeline of the AIS lifecycle.

Their proposed corrective action is a 'policy knot'—as a recurring loop between policy, design and operational practice that must hold across the full lifecycle of an AI-DSS.¹⁷ As the legal-analytical companion to this, Dorsey and Bo state:

[I]ncreased speed and scale in conflict made possible by the integration of AI-DSS, especially within the JTC [joint targeting cycle], can be a tradeoff with other considerations, such as the erosion or loss of human agency.¹⁸

While these studies focus on human agency, they also highlight the significance of human decision-making across the system lifecycle, motivating the need for a comprehensive discussion of *how* to preserve human decision-making in AI-DSS beyond the point of procedural use and by *what* means.¹⁹

The Institute of Electrical and Electronics Engineers Standards Association (IEEE SA) Research Group on Issues of Autonomy and AI in Defence Systems discusses the significance of technical testing, evaluation, verification and validation alongside human decision-making through the lens of human operators. While this could be characterised as a technical or policy challenge, we frame it also as a procedural challenge focused around the idea of human skills or human readiness level. Conn and Bode outline a need for

human readiness levels, modelled on the technical readiness level framework widely in use today. Australian defence policy articulates this need, noting that the policy takes a human-centric approach to consider impacts across all functions. The policy goes on to establish the 'accountable officer' framework within the capability lifecycle, uplifting accountability from an operational-only lens to include acquisition. Recent literature²⁰ highlights that acquisition can be deliberately structured as the implementation lever, with emerging defence acquisition reforms demonstrating progress systemically in Australia.

In this paper we introduce two critical military processes, the system lifecycle and the joint targeting process (JTP), highlighting how they intersect. We then discuss the complications of compromised technical integrity, focusing on three aspects central to AI-DSS in targeting operations: the speed and scale of target generation and nomination; the contention between accuracy and precision; and the explainability of the system outputs. We go on to explore the relationship between decision-making and time, and the impacts of AI-DSS on this relationship. In the final section we present recommendations for decision-makers across both the system lifecycle and the targeting operations.

The Cycles of Military Systems and Operations

Military operations are enabled and conducted via iterative and repetitive processes which have been established and refined over years, decades and centuries. These cycles facilitate procedural mechanisms for acquiring military systems and deploying them in operational contexts. There are two cycles we will focus on here, one at the acquisition level and the other at the operational level: the system lifecycle and the JTP respectively.

System Lifecycle

At the acquisition level, systems engineering is a standard framework employed by many states to procure and acquire military systems.^{21 22 23} Systems engineering is an interdisciplinary approach that enables the successful realisation, use and retirement of engineered systems.²⁴ This framework is built around the system lifecycle, a sum of phases and activities which bring a system into being, support and enable its use and then dispose of it once it can no longer serve its intended purpose.²⁵ The system lifecycle is a linear succession of phases that involves cyclical processes within each phase.

Human-made engineered systems come into being through 'purpose-driven human action'²⁶ across a system lifecycle. The significance of the human role in the system lifecycle process, particularly the weapon system lifecycle,²⁷ is critical to understanding the true impacts of AI-DSS used in targeting operations.

Various representations of system lifecycles exist. Within the context of this paper we will refer to the system lifecycle illustrated in Figure 1, adapted from our previous work.²⁸ We deliberately use an idealised end-to-end representation of a systems lifecycle, which has been adapted from the specific lifecycle in use in Defence and work from the IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems;²⁹ importantly, the lifecycle has been amended to capture necessary considerations for AI-enabled systems.

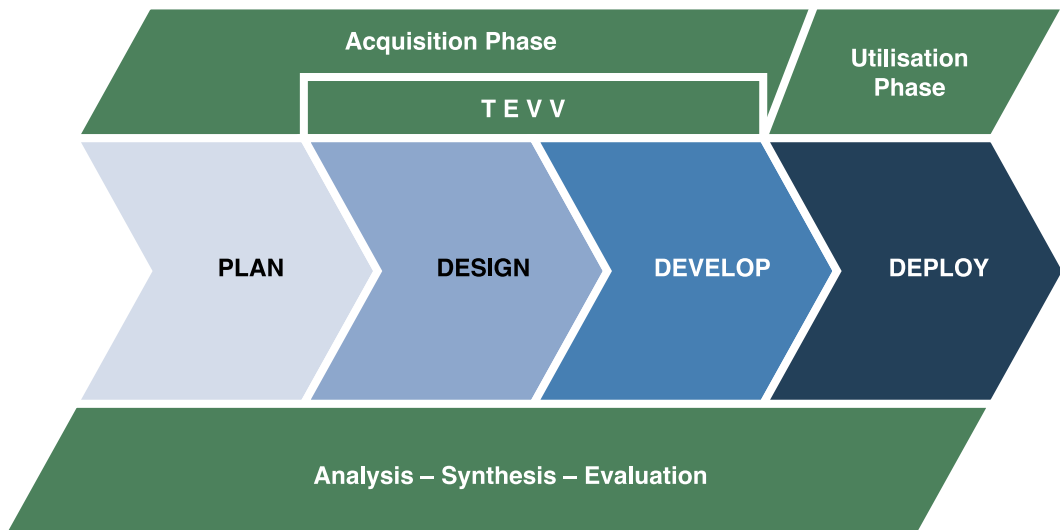


Figure 1. System lifecycle

This lifecycle involves four stages: three within the acquisition phase and the fourth within the utilisation phase of an engineered system. The four high-level lifecycles stages are:

1. **Plan.** The planning phase includes defining the scope and determining the objectives of a system.
2. **Design.** The design phase involves outlining specific requirements, architectures, functions, interfaces etc. of the system.
3. **Develop.** The development stage involves the actual building of the system, which includes testing, evaluation, verification, and validation (TEVV).
4. **Deploy.** The deployment stage involves integrating the system into operations, which includes maintenance throughout the life of the system.

The iterative and cyclical nature of systems engineering practices is seen in the repetitive application of analysis, synthesis and evaluation across all four lifecycle stages, which progress linearly over time.³⁰ Analysis resolves *what* is required and *why*, synthesis determines *how* it can be achieved, and evaluation assesses the trade-offs between the *what* and the *how*. A short note on terminology: acquisition and procurement are commonly used as synonyms; however, there is a distinction between the two terms. Procurement refers only to purchase, while acquisition encompasses the broader lifecycle process, including the act of procurement.³¹

One of the core phases of a system lifecycle is TEVV, which predominantly, but not exclusively, occurs during the design and development phases. Testing and evaluation involves assessing the performance, reliability and safety of a system. It involves systematically examining and validating these factors to ensure they meet specified requirements and perform as intended. Testing and evaluation has four major categories:

1. *Preview* test and evaluation (T&E) to assist analysis and refinement of major acquisition options for approval
2. *Development* T&E to support design and development efforts at the beginning acquisition stages
3. *Acceptance* T&E, which is formal acceptance testing conducted when the system is more mature
4. *Operational* T&E, which involves testing what the system would be like in operation, using operation-like conditions or scenarios.

Verification refers to the evaluation of whether a product, service or system complies with a regulation, requirement, specification or imposed condition. It is often an internal process. Validation refers to the assurance that a product, service or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability testing with external stakeholders. In essence, verification is the process of ensuring you have built the system right and validation is ensuring you have built the right system.

Targeting Cycle

The JTP is a cyclical and continuous activity consisting of six phases designed to support a commander in achieving an end state (detailed in Table 1).³² The JTP is the generic targeting methodology that has been used by the Australian military and developed against the US joint targeting cycle, which similarly consists of a six-phase process.³³

Table 1. Description of the six joint targeting process phases³⁴

Phase	Description
1	Commander's guidance Mission, objectives, intent and desired effects
2	Target development Intelligence direction, analysis, validation and target list management
3	Capabilities analysis Best available means to affect targets or target sets
4	Force application Assigning forces, weapons and/or other capabilities to targets
5	Execution Applying the force to realise objectives and desired effects
6	Assessment Effects assessment, battle damage assessment, weapons effectiveness assessment, collateral assessment and re-attack recommendations

Targeting refers to the process of ‘selecting and prioritizing targets and matching the appropriate response to them, considering operational requirements and capabilities’.³⁵ A target is defined as an object, person, organisation or geographic area which can be engaged or influenced in military operations to achieve a political end state.³⁶

It should be noted that the JTP represents one specific model for targeting; however, it is not a universally adopted model. What this model represents, more broadly, is the institutional approach to military targeting operations. While not universal, it is emblematic of military operational processes and how weapon systems are deployed in such systematic and cyclical processes.

Within the JTP, there are two categories of targeting: deliberate targeting, which involves planned targets; and dynamic targeting which involves ‘targets of opportunity’—targets not identified or selected in time to be included in deliberate targeting processes.³⁷ Both of these categories broadly follow the same six-phase targeting process, with the critical differentiator being time. Dynamic targeting deliberately compresses the target development phase with the intention of enabling engagement of ‘targets of opportunity’.

Similar to the system lifecycle, the JTP is an iterative and cyclical³⁸ process intended to manage decision-making, across multiple human actors and systems, to achieve a desired outcome. These two cycles intersect when an engineered military system, which has been acquired through a systems engineering process, is deployed and utilised in targeting operations which employ the JTP. The cycles of decision-making within the system lifecycle ultimately have flow-on effects on the targeting process. For AI-DSS used in targeting operations, those flow-on effects emerge through an erosion of technical integrity which compromises operational effectiveness and is compounded by temporally compressed decision-making during operations, as is discussed in the next section.

Through this work, we position the systems lifecycle (acquisition) as the upstream determinant of the JTP (operational)—two intrinsically connected processes. While human decision-making must be distributed across the total system, we argue that the most consequential system-level decisions are taken before the system is fielded.³⁹

Compromised Technical Integrity

Technical integrity refers to a system's fitness for service, safety and compliance with regulations.⁴⁰ Fit for service differs from fit for purpose: the former refers to meeting technical standards and operational requirements and the latter relates to accomplishing intended functions or outcomes. Fit for purpose is outcome focused, while fit for service is system focused and a pillar of technical integrity.

Technical integrity is a product of human decision-making and a reflection of human intervention in engineered systems. One of the predominant methods for enabling technical integrity is through TEVV, which largely, but not exclusively, occurs during the design and development stages of the system lifecycle. Accelerated development of AI technologies is achieved through accelerated, or in some cases overlooked,⁴¹ systems engineering practices, including TEVV, which underpin technical integrity. Technical regulations are a means of enforcing and assuring TEVV practices, and consequently are an additional pillar of technical integrity;⁴² however, the absence of such regulations specific to AI has created an exploitable gap which further compromises technical integrity.

The narrowing and potential omission of these critical processes can lead to an accumulative breakdown of decision-making over the course of the systems lifecycle of AI-DSS, culminating in the erosion of TEVV: the foundation of technical integrity. Technical integrity enables consistency and reliability of systems,⁴³ with the absence of these attributes resulting in inconsistent and unreliable systems. For example, the IDF's use of AI-DSS for targeting operations in Gaza has been the subject of substantial reporting on the performance limitations of these systems.⁴⁴ This point, more broadly, is structural and well established in the literature.⁴⁵ It also alludes to a broader provocation around acceptable risk thresholds for AI-enabled systems in military operations, an open question which is beyond the scope of this paper.

Offering a related critique of the argument in the absence of agreed AI risk thresholds, Khlaaf and Myers West state:

AI technologists have instead advocated for risk tolerances skewed by a purported AI arms race and speculative 'existential' risks, taking over the arbitration of risk determinations with life-or-death consequences, subverting democratic processes.⁴⁶

This perspective suggests that ongoing rhetoric about fabricated urgencies can mask irresponsible practices. We offer that this further motivates the need for deliberate treatment of the entire system lifecycle of AI-DSS, expanding beyond the weight of focus on operational-use phases.

The erosion of technical integrity has been identified as complicating three aspects central to AI-DSS use in targeting operations. These are the speed and scale of target generation and nomination; the contention between accuracy and precision; and the explainability of the system outputs. We suggest that all of these three points are manifestations of upstream decision-making which ultimately impact technical integrity and operational effectiveness, as we describe here.

Importantly, the Israel–Gaza war, the Russia–Ukraine war and Operation Epic Fury demonstrate, at present, the most in-depth examples of the use of AI-DSS in targeting operations. For these conflicts—in contrast to previous ones—much of the information on the specifics of use and associated limitations is presented and available in public reporting, as opposed to academic literature; however, many aspects of reported claims are substantiated by technical publications in academic discourse, which we present alongside our argument.⁴⁷

Speed and Scale of Target Generation and Nomination

Multiple investigative reports into the IDF's use of AI-DSS in targeting operations document the increased speed and scale of target generation and nomination afforded by the use of these systems. Retired Lt Gen Aviv Kohavi, IDF Chief of Staff until 2023, described The Gospel as capable of producing 100 targets per day (over 36,500 annually), against approximately 50 per year for human analyst—a rate increase exceeding two orders of magnitude.^{48 49 50 51} It has been reported, although also contested, that human analysts were making decisions regarding a single target within 20 seconds, prior to authorisation. This is consistent with the technical capacity of AI-DSS, which are capable of analysing very large datasets within compressed time frames.⁵² Target development, phase two of the JTP, involves intelligence direction, analysis, validation and target list management—tasks traditionally distributed across multiple human analysts. The deployment of AI-DSS in the targeting process compresses these tasks temporally and procedurally, as well as informationally.

While speed is a factor in targeting operations, particularly for dynamic targeting, speed at the expense of operational effectiveness presents a paradoxical challenge. A system is deemed fit for purpose when it accomplishes or supports the accomplishment of intended functions or outcomes. In the case of targeting operations, a system deployed in the JTP will be deemed fit for purpose if targets are correctly identified and engaged in a manner deemed operationally effective; here we take that to mean compliant with IHL principles of distinction and proportionality, as we outlined previously in this paper.

For AI-DSS, the scaled and accelerated identification of targets may allude to a fit-for-purpose system; however, the accuracy of those identifications and the precision of their engagement lends a different perspective. A system can be fit for purpose but not fit for service, an underpinning principle of technical integrity. AI-DSS used in targeting operations are an example of a type of system which could be argued to be fit for purpose but not fit for service.

The erosion of technical integrity which would deem a system fit for service can be the result of the narrowing and omission of critical military processes, leading to an accumulative degradation of decision-making upstream of the point of operational use. The acquisition phase of the system lifecycle—plan, design, develop—traditionally consists of decision-making windows which allow for considerations around objectives of a system, identifying requirements, architectures and functions, legal obligations, and TEVV, among others. The shifting military acquisition landscape has temporally and procedurally compressed these opportunities for intervention. This may be why we are seeing the deployment of AI-DSS in targeting operations which are fit for purpose but not fit for service.

The erosion of technical integrity of AI-DSS used in targeting operations paradoxically complicates the allure of speed and scale of target generation and nomination because the claims of efficiency and effectiveness are contradicted by the reality of operational outcomes. If these systems were as efficient and effective as claimed, it would be logical to assume civilian casualties from targeting operations involving AI-DSS would be minimal, in line with the definition of operational effectiveness presented herein. However, reporting on the civilian death toll in Gaza points to a disproportionately high number of non-combatant casualties.^{53 54 55}

The use of AI-DSS in targeting operations is not the sole factor contributing to limited operational effectiveness. This outcome is also a product of deploying systems which lack technical integrity into targeting operations, conducted without deliberate, considered and timely decision-making. We note that, specifically regarding the IDF, speed and scale have historically been prioritised in their intelligence gathering and targeting activities.^{56 57}

The degradation of upstream decision-making begins to bleed into operational decision-making through the encouragement of temporally and procedurally compressed decision-making practices during targeting operations. The pursuit of speed and scale begins upstream of operations, during the acquisition phase of the system lifecycle, and is carried forward into operational practices in the JTP.

On the first morning of Operation Epic Fury, US forces struck the Shajareh Tayyebbeh primary school in Minab, in southern Iran, killing approximately 156 people, including 120 children, most of them girls aged between seven and twelve.^{58 59} A preliminary US military

investigation determined US responsibility for the strikes. As widely reported, two AI systems were used: Palantir's Maven Smart System for targeting and querying underlying databases, and Anthropic's Claude large language model in a decision-support capacity.⁶⁰ The Commander of US Central Command, US Navy Admiral Brad Cooper, has stated on public record:

these systems help us sift through vast amounts of data in seconds so our leaders can cut through the noise and make smarter decisions faster than the enemy can react. Humans will always make final decisions on what to shoot and what not to shoot and when to shoot.⁶¹

In 2016, the building had been separated from an adjacent Islamic Revolutionary Guard Corps compound and converted into a school. The change was noted to be visible on both satellite imagery and local business directories.⁶² An investigation into this incident found that an analyst had recorded their observation of the change in 2019; however, the analyst's record was not entered into a digital tool connected to a US military official targeting database.⁶³ This critical piece of information was therefore missing from the targeting process. A subsequent investigation found repeated decision-making failures over several years, pointing to broader systemic issues. Our focus here is on the disconnect between upstream decision-making and operational outcomes.

The design and development phases of the system lifecycle include activities around implementation compatibility: determining how a system can be robustly and effectively integrated alongside other systems.^{64 65} The investigation into the Minab school bombing revealed a disconnect and a level of incompatibility among the many systems, both technical and operational, utilised in military targeting operations, highlighting the true complexity of system implementation—a complexity which requires its own inquiry that is beyond the scope of this paper. As previously noted, AI systems are not deployed in a vacuum; rather they are implemented into existing military systems and procedures, some of which they complement and others which they conflict with.

The preliminary findings indicate an accumulation of failures across the system lifecycle, of which TEVV practices and consideration of implementation compatibility in the acquisition phase could possibly have minimised the likelihood. A mitigation mechanism such as mandated data quality assurance practices may have emerged during the TEVV process, when limitations of a system are identified, through testing and evaluation, and mitigations are developed to compensate for them. This is how consistency and reliability—that is, technical integrity—can be assured. As processes are being narrowed, or in some cases omitted, the opportunities for human intervention, in the form of decision-making, are also being constrained. Errors can propagate into the targeting process, in which a verification of the target's status would have been likely to result in a revised assessment.

Table 2 summarises open-source reporting of AI-DSS functions deployed in Operation Epic Fury, and the role of these technologies in decision support. As with the use of AI across a wide range of applications, both civilian and military, we observe that a single, monolithic AI-DSS was not used. Rather, a series of discrete technologies were used to support specific decision outcomes. From available information, it is unclear how decision inputs were handled and the ways in which errors manifested through decision-making processes. However, the accumulation of errors across successive decisions does highlight the implications of accumulative decision failure, and how this can rapidly propagate to lethal outcomes.

Table 2. Analysis of AI-DSS roles in the Shajareh Tayyebbeh strike^{66 67 68}

Decision input	AI-DSS function	Technology
Sensing	Integration of satellite imagery, signals intelligence and uncrewed system sensor data	Maven Smart System, machine learning vision models and machine learning sensor models
Target detection and classification	Object recognition and detection confidence scoring	Maven Smart System
Intelligence contextualisation	Template matching (algorithmic detections correlated to military facility records)	Human-curated Defense Intelligence Agency database with legacy data (not updated post 2016 changes)
Course-of-action development and recommendation	Targeting prioritisation	Maven Smart System
Intelligence curation	Search, summarisation and synthesis of intelligence reporting	Large language model, reportedly Anthropic's Claude

The emergence of narrowing and omission of critical processes through temporally and procedurally compressed decision-making for AI-DSS in the acquisition phase is likely to be a key contributing factor in technical integrity erosion, which manifests during the deployment of AI systems into targeting operations, leading to compromised operational effectiveness. It is this accumulation of degraded decision-making across the entire system lifecycle—from acquisition through to utilisation—which leads to outcomes such as the example provided.

Three months after Russia's invasion of Ukraine, Palantir Technologies chief executive Alex Karp arrived in Kyiv to meet President Zelensky, offering to deploy the company's data and artificial intelligence software in support of Ukraine's defence.⁶⁹ Palantir subsequently entered agreements with at least six Ukrainian agencies, including the Ministry of Defence, the Ministry of Digital Transformation, the Ministry of Economy, and the Ministry of Education.⁷⁰ The MetaConstellation platform, now operational,

tasks commercial satellite imagery from Maxar Technologies, Planet Labs and Capella Space; integrates signals intelligence; and produces target packages which, in Karp's own assessment, render Palantir software 'responsible for most of the targeting in Ukraine'.⁷¹ The specifics of the agreements remain confidential.⁷²

This case highlights the emerging commercialisation of military decision-making processes through acquisition, with essential operational and intelligence decisions potentially outsourced to commercial entities. While the need for procurement as implementation here is evident, greater risks may be yet to manifest downstream. The Center for Security and Emerging Technology's Margarita Konaev has framed the risk as follows:

Most companies operating in Ukraine right now say they align with U.S. national-security goals—but what happens when they don't? What happens the day after?⁷³

We have described technical integrity as a means of enabling consistency and reliability in systems. In the case of AI-DSS, these systems may be capable of scaling and accelerating target identification; however, issues with consistency and reliability emerge in the accuracy and precision of those targets and the method of their engagement.

Contention Between Accuracy and Precision

Accuracy and precision are two terms which mean different things at a technical level. In computer science, precision refers to consistency evident through a lack of, or minimal, variance; whereas accuracy refers to the correctness of outputs.⁷⁴ For example, if an AI-DSS *correctly* identified or classified targets, it would be deemed an accurate system, and if it *consistently* produced the same identification and classification results for individual targets when prompted multiple times, it would be deemed a precise system. An AI-DSS which produces incorrect target identifications and classifications would be deemed inaccurate; however, if it produced the same output multiple times, despite the output being incorrect, then it would be considered precise. It is for this reason that the distinction between accuracy and precision is important to delineate. The characteristic of precision for software systems is not an indication of accuracy or correctness, and while it may make a system fit for purpose, it does not necessarily make it fit for service.

Precision and accuracy are technical factors which are commonly assessed during TEVV and contribute to the consistency and reliability of a system. What is challenging about AI-DSS used in targeting operations is the dependency on non-technical factors for ensuring technical accuracy. Some of these factors can be captured during the acquisition phase, while others require tailored decision-making at the operational level.

Complexity is an emergent property of AI-DSS,⁷⁵ and AI-enabled systems more broadly, wherein certain behaviours or outputs only emerge at the system of systems level, in which there are interconnections and interactions with other elements.⁷⁶ One of those elements is data. All AI-enabled systems are data-dependent systems which produce outputs based on training datasets. It is almost impossible to curate a dataset, for any industry or context, which captures every aspect of an operating environment.⁷⁷ This is especially true for military operating environments, which are inherently dynamic and often unpredictable. An AI-DSS may be considered precise and accurate prior to deployment, based on its performance against a bounded use case represented in its training dataset, but inaccurate once deployed in a context which exceeds the bounds of its training dataset.

For AI-DSS, or AI systems more broadly, accuracy is neither static nor definitive. This is why a system may be deemed fit for service at the point of deployment but its technical integrity may erode once it is deployed in targeting operations in which accuracy, an operational requirement of the system, becomes compromised. Here, the erosion of technical integrity did not begin at the point of deployment; rather, it began upstream of the point of use as a result of narrowed or omitted TEVV processes, the implication of which was amplified when the system was deployed in the targeting process.

One of the major categories of testing and evaluation is operational testing, which involves testing what the system would be like in operation using operation-like conditions or scenarios. While it is true that the dynamic and unpredictable nature of a military operating environment is almost impossible to completely capture in a dataset, the limitations that come with this reality must be mitigated through both technical interventions (technical controls) and operational practices (procedural controls). One example of an operational practice to mitigate accuracy limitations is verification requirements of identified targets during targeting operations that extend beyond a performative 'rubber stamp'. This may include integrity checks of information—integrity in this context refers to how much the information can be relied upon (trusted). Mitigation strategies and redundancy measures, at both the system and operational levels, are standard outputs of the TEVV process. Because the TEVV process is being procedurally and temporally compressed with AI-enabled systems, the foresight that traditionally comes with these practices is now evidently missing. This raises questions around the extent to which these systems are truly fit for purpose.

Verification is the process of evaluating whether a system is fit for service against regulations and technical requirements and specifications, and validation determines whether it is fit for purpose against stakeholder needs. The former underpins technical integrity, while the latter supports customer needs and expectations. In this paper we have argued that upstream decision-making erodes technical integrity across the system lifecycle, and this erosion of technical integrity is evidently complicating fitness of purpose

during system operation. Different decision-making upstream of the point of use can minimise, but not eliminate, the potential for system inaccuracies through technical interventions where feasible, and can mitigate the impacts of inevitable inaccuracies during operation through operational mitigation and redundancy measures. These opportunities for human intervention, through decision-making during the acquisition phase, can be narrowed and/or omitted, with consequences that manifest during the operational phase when these systems are deployed in targeting operations. The result is systems which are neither fit for service nor fit for purpose.

When implemented in targeting operations, such as the JTP, the accumulative degradation of decision-making is compounded by the absence of operational directives specific to the limitations of the system, which traditionally help to compensate for these limitations. The systemic and procedural nature of the JTP, be it in deliberate or dynamic targeting, may give the semblance of structure to targeting operations which employ AI-DSS; however, the reality of eroded technical integrity is the absence of consistency and reliability, which is evident in the contention between precision and accuracy.

This scenario highlights the intrinsic connection between the acquisition phase and the targeting process; how the contention between precision and accuracy at a technical level complicates the classification of systems as fit for purpose, fit for service, or neither; and the cascading impacts of that on operational effectiveness.

Explainability of System Outputs

The intended purpose of AI-DSS is to support, rather than replace, human decision-making.^{78 79} Ultimately, the decision to engage an identified target sits with a human, regardless of the system or systems deployed within the targeting process to support that decision. A critical part of decision-making is analysing and interpreting information, a task which has been absorbed by AI-DSS because of their capacity to analyse large amounts of information in rapid timeframes. While the analysis and interpretation of information may be delegated to AI-DSS, the requirement for interpreting system outputs to inform decision-making remains with humans.

It is difficult to interpret system outputs when the logic for how those outputs were produced is not clear or transparent.⁸⁰ The outputs of AI-DSS are not readily interpretable because the computational process for achieving those outputs is too complex, which is why these systems are labelled 'black box' systems.⁸¹ This is where the contention between precision and accuracy further complicates operational effectiveness.

Misinterpretation of information can result in inaccurate perceptions or understandings of a military operation, and imprecise or inaccurate AI-DSS present greater opportunities for misinterpretation, particularly when the opacity of these systems makes interrogation of their outputs difficult. If an AI-DSS is precise but inaccurate, it may produce the same

incorrect output multiple times, leading to the assumption that the output is correct because the system consistently produces it. In the absence of explainability of the system output, detecting that inaccuracy may not be straightforward.

In addition to consistency and reliability, what technical integrity also enables is trust. In the context of this paper, trust is defined as confidence in the reliability of a system when used in the intended operation of use, a definition introduced in previous work by the authors.⁸² While the topic of trust in the context of AI-enabled systems warrants its own paper, here we highlight one of the important elements of trust: trust calibration. Over-trust or under-trust in AI-enabled systems leads to ineffective outcomes, with a calibrated level of trust being the ideal middle ground.⁸³ When the contention between precision and accuracy is presented against an opaque AI-DSS in which the system outputs are not readily explainable, the result is miscalibrated trust. Over-trust leads to overconfidence in system outputs under the assumption the system is always correct,⁸⁴ and under-trust leads to underutilisation of system capabilities through misuse or dismissal of system outputs.⁸⁵

Explainability of system outputs is a means of supporting calibrated levels of trust in AI-DSS, a necessary requirement when these systems are deployed in targeting operations. Over-trust in system outputs can encourage a de-emphasis on the need for human verification of targets, leading to humans acting as a superficial 'rubber stamp' during targeting operations.⁸⁶ Under-trust in system outputs may lead to the dismissal of identified targets or to missed 'targets of opportunity' in dynamic targeting operations.

Explainability of system outputs can be identified as a requirement of AI-DSS during the acquisition phase—noting that expectations around what that explainability looks like must align with technical feasibility. Additionally, robust TEVV practices upstream of the point of use would soften the contention between precision and accuracy and would enforce consistency and reliability in the system in support of technical integrity, resulting in systems which have a greater likelihood of promoting calibrated levels of trust during operation.

The Relationship Between Decision-Making and Time

There is a strong relationship between decision-making and time. The use of AI in the military domain has impacted this relationship by compressing both decision-making timeframes and the opportunities to make decisions. For the system lifecycle, decision-making is static in nature, while the targeting process involves dynamic decision-making. Static decision-making treats the decision task and process as fixed in time, while dynamic decision-making takes into account either the duration of decision-making, the optimal time to make a decision, or changes in the decision structure as a function

of time.⁸⁷ The system lifecycle involves static decision-making because the sequence of decisions across the linearly progressing lifecycle phases are not subject to constant change, whereas targeting operations involve dynamic decisions resulting from the intrinsically dynamic nature of military operating environments.

The ongoing AI arms race has incited a ‘fabricated fear of falling behind’,⁸⁸ which has consequently accelerated the acquisition of AI-enabled military systems. The pace of planning, designing and developing engineered systems in the private technology sector has been condensed to promote technology adoption at unprecedented speed and scale. This compressed process narrows the opportunities for human intervention and critical decision-making windows. AI systems further exploit the factors of speed and scale through their capacity to analyse large datasets in rapid timeframes. When implemented into the targeting processes, this capacity for narrow decision-making timeframes shifts the relationship between decisions and time within targeting operations.

Compromised technical integrity is compounded by temporally compressed, high-stakes decision-making during targeting operations. Although contested, reporting on the IDF’s use of AI-DSS indicates that approximately 20 seconds is devoted to reviewing AI-DSS generated targets before engagement authorisations.⁸⁹ This limited timeframe calls into question the purpose of humans in this evolved targeting process. While dynamic targeting is designed to be at a higher tempo in comparison to deliberate targeting, the targeting process, as outlined in Table 1, still requires a minimum viable time to be completed. Accelerated operational decision-making not only has the potential to result in ineffective operational outcomes; it also disrupts broader geopolitical tensions as ‘speed optimised past the pace of human judgment is itself an escalation mechanism’⁹⁰—an escalation mechanism hidden under the veneer of a data-driven approach.⁹¹

Military operations are often framed as time sensitive or ‘high tempo’;⁹² hence the speed characterisation of AI-DSS fits within this framing as a positive intervention. Military operations involve cycles of decision-making, as is outlined in the previous section, and decision-making requires time, a requirement which challenges the purported benefit of AI-DSS. When these two time-dependent constructs are combined, balancing between pursuing speed and preserving time for decision-making becomes a difficult and precarious task. In the case of current deployments of AI-DSS in targeting operations, the scales are tipped in favour of accelerated decision-making, leading to compromised operational effectiveness, as was evident in the Minab school bombing.

Balancing decision-making and time is ultimately a trade-off between the benefits and shortcomings on either side. The inherent safety-critical nature of military operations adds an additional layer of complexity in the form of catastrophic outcomes when things go wrong. The nature of safety-critical operations, such as targeting operations, demand time and consideration in decision-making. However, for targeting operations, there may

be circumstances in which that time and consideration cannot be afforded. The issue we are seeing with AI-DSS used in targeting operations in current conflicts is the exploitation of operational time sensitivity which encourages accelerated decision-making as a standard practice. Accelerated decision-making aided by AI-DSS that lack appropriate human oversight and thorough technical integrity may be exposing military forces to inconsistent and unreliable systems informing decisions of life and death.

Recommendations

Recommendation 1: Technical integrity should be a mandatory requirement of AI-DSS intended for use in targeting operations, with evidence of ‘fitness for service’ demonstrated upstream of the point of use.

In this work we have argued that AI-DSS employed for targeting operations are at risk of being fit for purpose but not fit for service. Goussac and Boulanin frame procurement as the mechanism that serves to implement political commitments and legal obligations. Contracting within the acquisition stage is highlighted as:

a critical opportunity for implementing principles of responsible behaviour because it determines what the supplier must deliver and prove (including claims about the results of testing and assurance processes), who is accountable for certain risks and failures, and what obligations and requirements are borne by the parties.

While the Policy Settings for Responsible Use of AI in Defence establish a values-based architecture, we advocate for further technical integrity depth to meaningfully operationalise the policy. This may include the development of fit-for-service evidence, such as contextual TEVV under realistic conditions to draw out system friction and fracture points, draw out failure modes and establish robust mitigation strategies. The caution of Conn and Bode that ‘if the use of the system seems to turn human users into button pushers and therefore risks only securing nominal human control ... this needs to be flagged in earlier stages during testing’ applies directly: fitness-for-purpose performance is not a substitute for evidence of fitness for service.

Recommendation 2: Speed of operational decision-making should be assessed against operational effectiveness, which should include IHL obligations.

High-tempo operations do not always lead to operationally effective outcomes. Accordingly, the potential advantages of accelerated operational decision-making should be assessed against operational effectiveness. In this paper, we frame operational effectiveness as verification leading to protection, including the safety of non-combatants

and combatants, consistent with the IHL principles of distinction (distinguishing between combatants and non-combatants) and proportionality (ensuring that disproportionate impacts on civilians during military operations are minimised).

Recommendation 3: AI-DSS operational effectiveness for targeting operations should be quantitatively verified against established benchmarks and validated against the commander's intent prior to deployment.

A complete operational effectiveness assessment for AI-DSS in targeting operations should evaluate performance against four dimensions: mission achievement, force preservation, operational tempo and IHL compliance. Validation against a commander's intent provides the standard against which trade-offs between these dimensions are resolved, and retains human command and operator engagement throughout critical phases of the targeting cycle. Consideration should be given to the design and interface of AI-DSS to support human engagement, negating potentially irreversible 'automatic acceptance' of outputs.⁹³

Conclusion

This paper has examined the lifecycle of AI-DSS deployed in military targeting, exploring failure modes evident at the point of use resulting from upstream decisions in acquisition and capability development. The omission of critical military processes results in an accumulative degradation of decision-making across the lifecycle, eroding technical integrity of the system and compromising operational effectiveness in the targeting operations for which such systems are deployed. We have explored the intersection of acquisition and operational processes through a use-case based approach to highlight potential pitfalls and challenges that arise from the integration of AI-DSS. The forward-looking question is not whether AI-DSS will be deployed in targeting; these systems are already deployed. This paper highlights the compounding impacts of narrowed and omitted critical military processes across the system lifecycle and how these manifest when systems are utilised in military operations, specifically targeting operations. The question is whether the institutional governance architecture of acquisition, command accountability and operational TEVV will adapt to evolving military procurement processes.

About the Authors

Dr Zena Assaad is an Associate Professor in Engineering at the Australian National University (ANU) and the Deputy Director of the ANU Defence Institute. She is currently a fellow with the UN Institute for Disarmament Research (UNIDIR) leading research on the decommissioning of military AI systems. She has held fellowships with the Australian Army Research Centre and Trusted Autonomous Systems. Her research explores the safety of critical and complex technologies in safety-critical domains, with a focus on systems engineering practices. She is currently focused on topics related to autonomy and AI in military systems. Dr Assaad is a former member of the expert advisory group for the global commission on responsible AI in the military domain (GC REAIM). She is currently an Ambassador for the Responsible by Design initiative of Article 36 Legal. She also serves on the independent advisory board on Legal Reviews as a technical expert. She is also the host of the Responsible Bytes podcast.

Lieutenant Colonel Dr Adam Hepworth is presently the Australian Army Liaison Officer to the United States Army Transformation and Training Command, based in Austin, Texas. Previously, Adam was Director of the Army Robotic and Autonomous Systems Implementation and Coordination Office (RICO). He is presently a Visiting Fellow at the University of New South Wales and an Advisory Board Member at the Responsible by Design Institute. Adam has formerly been an Expert Member on the Global Commission for Responsible AI in the Military Domain (GC REAIM), an expert advisor to the United Nations Institute for Disarmament Research (UNIDIR) on land warfare and autonomous systems, and an expert advisor to the International Committee of the Red Cross (ICRC) on Artificial Intelligence Decision Support Systems (AI-DSS). Adam holds a Bachelor of Science in Mathematics from the University of New South Wales, a Master of Logistics and Supply Chain Management from the University of South Australia, a Graduate Diploma in Scientific Computation and a Master of Science in Operations Research from the United States Naval Postgraduate School, and a Doctor of Philosophy in Computer Science from the University of New South Wales. He is a Graduate of the Australian Institute of Company Directors (GAICD).

ENDNOTES

- 1 Michael Biesecker, Sam Mednick and Garance Burke, 'As Israel Uses US-Made AI Models in War, Concerns Arise About Tech's Role in Who Lives and Who Dies', *AP News*, 18 February 2025.
- 2 Alexander Blanchard and Laura Brunn, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses* (Stockholm International Peace Research Institute, 2025).
- 3 Katrina Manson, 'US Military Relying on AI as Tool to Speed Iran Operations', *Bloomberg*, 5 March 2026.
- 4 'Anthropic's AI Tool Claude Central to US Campaign in Iran, Amid a Bitter Feud', *The Washington Post*, 4 March 2026.
- 5 AD Larson and CC Hayes, 'An Assessment of Weasel: A Decision Support System to Assist in Military Planning', *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* 49, no. 3 (2005): 287–291.
- 6 'Accelerating Decision Making: National Geospatial-Intelligence Agency at AIPCon 5' (video), Palantir YouTube channel, 19 September 2024, at: www.youtube.com/watch?v=XzKnUt6NAbw.
- 7 Anna Nadibaidze, Ingild Bode and Qiaochu Zhang, *AI in Military Decision Support Systems: A Review of Developments and Debates* (Odense: Center for War Studies, 2024).
- 8 Human Rights Watch, 'Questions and Answers: Israeli Military's Use of Digital Tools in Gaza', *Human Rights Watch* (website), at: www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-digital-tools-gaza.
- 9 Gaby Tejeda, 'AI Integration in Operation Epic Fury and Cascading Effects', *The Soufan Center* (website), 3 March 2026, at: <https://thesoufancenter.org/intelbrief-2026-march-3>.
- 10 Elizabeth Dwoskin, 'How Israel Built an "AI Factory" for War, Use in Gaza', *The Washington Post*, 29 December 2024.
- 11 David Hollingworth, 'Machine War: How AI and Other Technologies Are Shaping the Fighting in Gaza', *Defence Connect*, at: www.defenceconnect.com.au/joint-capabilities/16182-machine-war-how-ai-and-other-technologies-are-shaping-the-fighting-in-gaza.
- 12 'Reported Impact Snapshot I Gaza Strip (20 August 2025)', *United Nations Office for the Coordination of Humanitarian Affairs* (website), at: www.ochaopt.org/content/reported-impact-snapshot-gaza-strip-20-august-2025.
- 13 Noa Yachot, "'Data Is Control': What We Learned from a Year Investigating the Israeli Military's Ties to Big Tech", *The Guardian*, 31 December 2025.
- 14 Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence* (Canberra: Commonwealth of Australia, 2025).
- 15 Netta Goussac and Vincent Boulanin, *Responsible Procurement of Military Artificial Intelligence* (Stockholm: Stockholm International Peace Research Institute, 2026).
- 16 Conn and Bode define the term AIS as 'artificial intelligence system'.
- 17 A Conn and I Bode, 'Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems', *Ethics and Information Technology* 27, no. 51 (2025).
- 18 Jessica Dorsey and Marta Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human(e) Dimension', *International Law Studies* 107 (2026): 184–229.
- 19 A recent study by the Boston Consulting Group, published in the *Harvard Business Review*, outlines implications of 'AI brain fry' (AI mental fatigue). The report highlights the capacity limit of human biological brains, noting that the strongest predictor of mental fatigue was the level of oversight required for the AI agent (i.e. greater intimate oversight leads to greater levels of fatigue). See Julie Bedard, Matthew Kropp, Megan Hsu, Olivia T Karaman, Jason Hawes and Gabriella Rosen Kellerman, 'When Using AI Leads to "Brain Fry"', *Harvard Business Review*, 5 March 2026.
- 20 Goussac and Boulanin, *Responsible Procurement*.
- 21 Alexandre Verlaine, 'No-Capability Defence Acquisition: A Literature Review on Policy and Practice of Military Acquisition Within NATO and the EU', *International Journal of Procurement Management* 17, no. 4 (2023): 443–487.

- 22 Desmond Ball, 'Arms and Affluence: Military Acquisitions in the Asia-Pacific Region', *International Security* 18, no. 3 (1993): 78–112.
- 23 W Perry, 'US Military Acquisition Policy', *Comparative Strategy* 13, no. 1 (1994): 19–24.
- 24 Cecilia Haskins, Kevin Forsberg and Michael Krueger, *Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities*, version 3.1 (International Council on Systems Engineering, 2007).
- 25 I Faulconbridge and MJ Ryan, *Systems Engineering Practice* (Argos Press, 2014).
- 26 BS Blanchard and WJ Fabrycky, *Instructor's Guide to Problem Solutions for Systems Engineering and Analysis*, fifth edition (Pearson, 2011).
- 27 JR Nelson, P Konoske Dey, MR Fiorello, JR Gebman, GK Smith and A Sweetland, *A Weapon-System Life-Cycle Overview: The A-7D Experience*, R-1452-PR (RAND, 1974).
- 28 Z Assaad and A Hepworth, *A Systems Engineering Lifecycle Approach to Responsible AI*, (The Hague Centre for Strategic Studies, 2025).
- 29 S Allik et al., *A Framework for Human Decision Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (New York: IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, 2024).
- 30 Faulconbridge and Ryan, *Systems Engineering Practice*.
- 31 RP Smith, *Defence Acquisition and Procurement: How (Not) to Buy Weapons* (Cambridge University Press, 2022).
- 32 Australian Defence Force Warfare Training Centre, *Operations Series ADDP 3.14 Targeting* (Australian Defence Doctrine Publications, Defence Publishing Service, 2009)
- 33 Mitch Ferry, 'F3EA: A Targeting Paradigm for Contemporary Warfare', *Australian Army Journal* 10, no. 1 (2013).
- 34 Australian Defence Force Warfare Training Centre, *Operations Series ADDP 3.14*.
- 35 U.S. Department of Defense, *DOD Dictionary of Military and Associated Terms* 232, Joint Publication 1-02 (Department of Defense, 2017).
- 36 *Allied Joint Doctrine for Joint Targeting*, AJP-3.9 (NATO, 2008).
- 37 Australian Defence Force Warfare Centre, *Operations Series ADDP 3.14*.
- 38 Dorsey and Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle', pp. 184–229.
- 39 Conn and Bode, 'Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems'; Allik et al., *A Framework for Human Decision Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.
- 40 MT Edwards, *Measuring the Technical Integrity of a Complex Engineered System* (University of South Australia, 2014).
- 41 Heidy Khlaaf and Sarah Myers West, 'Safety Co-Option and Compromised National Security: The Self-Fulfilling Prophecy of Weakened AI Risk Thresholds', *arXiv Computers and Society*, arxiv:2504.15088 (2025).
- 42 Michael Edwards, '7.3.1 The Impact of Technical Regulation on the Technical Integrity of Complex Engineered Systems', *INCOSE International Symposium* 19, no. 1 (2014): 1139–1153.
- 43 Z Assaad, 'A Framework for Safely Scaling Multi-Agent Teams', *Australian Army Journal* 21, no. 2 (2025): 132–158.
- 44 Dorsey and Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle', p.187.
- 45 For instance, see Allik et al., *A Framework for Human Decision Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*, p. 18; Conn and Bode, 'Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems'; Dorsey and Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle', p. 226; Lewis, Michael. 2013. "Human Interaction with Multiple Remote Robots." *Reviews of Human Factors and Ergonomics* 9 (1): 131–74. <https://doi.org/10.1177/1557234X13506688>.
- 46 Khlaaf and Myers West, 'Safety Co-Option and Compromised National Security'.

- 47 We acknowledge that the information environment surrounding this conflict is heavily contaminated with AI-generated claims (fabrications); hence this paper attributes contested claims to identified sources rather than as assertions.
- 48 Yuval Abraham, “‘Lavender’: The AI Machine Directing Israel’s Bombing Spree in Gaza”, *+972 Magazine*, 3 April 2024.
- 49 Lauren Gould, Linde Arentze and Marijn Hoijtink, ‘Gaza War: Artificial Intelligence Is Changing the Speed of Targeting and Scale of Civilian Harm in Unprecedented Ways’, *The Conversation*, 24 April 2024.
- 50 Elizabeth Dwoskin, ‘How Israel Built an “AI Factory” for War, Use in Gaza’, *The Washington Post*, 29 December 2024.
- 51 Harry Davies, Bethan McKernan and Dan Sabbagh, “‘The Gospel’: How Israel Uses AI to Select Bombing Targets in Gaza”, *The Guardian*, 1 December 2023.
- 52 Z Assaad and E Williams, ‘Technology and Tactics: The Intersection of Safety, AI, and the Resort to Force’, *Cambridge Forum on AI: Law and Governance* 1, no. 49 (2025).
- 53 Neta Crawford, ‘Gaza: Civilian Death Toll Outpaces Other Modern Wars’, *The Conversation*, 27 August 2025.
- 54 ‘Reported Impact Snapshot: Gaza Strip (20 August 2025)’, *United Nations Office for the Coordination of Humanitarian Affairs* (website), at: www.ochaopt.org/content/reported-impact-snapshot-gaza-strip-20-august-2025.
- 55 Ozer Khalid, ‘Gaza: Ground Zero for AI Warfare’, *Criterion Quarterly* 19, no. 4 (2025).
- 56 Yachot, “‘Data Is Control’”.
- 57 ‘Israel and the Occupied Palestinian Territory’, *Amnesty International* (website), at: www.amnesty.org/en/location/middle-east-and-north-africa/middle-east/israel-and-the-occupied-palestinian-territory/report-israel-and-the-occupied-palestinian-territory.
- 58 Kevin T Baker, ‘AI Got the Blame for the Iran School Bombing. The Truth Is Far More Worrying’, *The Guardian*, 26 March 2026.
- 59 ‘USA/Iran: Those Responsible for Deadly and Unlawful US Strike on School That Killed Over 100 Children Must Be Held Accountable’, *Amnesty International* (website), 16 March 2026, at: www.amnesty.org/en/latest/news/2026/03/usa-iran-those-responsible-for-deadly-and-unlawful-us-strike-on-school-that-killed-over-100-children-must-be-held-accountable.
- 60 Baker, ‘AI Got the Blame for the Iran School Bombing’.
- 61 Jon Harper, ‘Centcom Commander Touts Use of AI in Fight Against Iran During Operation Epic Fury’, *DefenseScoop*, at: <https://defensescoop.com/2026/03/11/us-military-using-ai-against-iran-operation-epic-fury-adm-cooper>.
- 62 Baker, ‘AI Got the Blame for the Iran School Bombing’.
- 63 Katrina Manson, ‘An Analyst’s Missed Remark Surfaced in Deadly Iran School Strike Probe’, *Bloomberg*, 26 June 2026.
- 64 J Floch, C Carrez, P Cieslak, M Roj, RT Sanders and MM Shiaa, ‘A Comprehensive Engineering Framework for Guaranteeing Component Compatibility’, *Journal of Systems and Software* 83, no. 10 (2010): 1759–1779.
- 65 Mikela Chatzimichailidou, Tim Whitcher and Nikola Suzic, ‘Complementarity and Compatibility of Systems Integration and Building Information Management’, *IEEE Systems Journal* 18, no. 2 (2024): 1198–1207.
- 66 Baker, ‘AI Got the Blame for the Iran School Bombing’.
- 67 Tejada, ‘AI Integration in Operation Epic Fury and Cascading Effects’.
- 68 Manson, ‘US Military Relying on AI as Tool to Speed Iran Operations’.
- 69 V Bergengruen, ‘How Tech Giants Turned Ukraine Into an AI War Lab’, *Time*, 9 February 2024.
- 70 Ibid.
- 71 Ibid.
- 72 Y Mysyshyn, *Advanced Technologies in the War in Ukraine: Risks for Democracy and Human Rights* (Washington DC: German Marshall Fund of the United States, 2024).
- 73 Ibid.

- 74 Ronald F Boisvert, Ronald Cools and Bo Einarsson, 'Assessment of Accuracy and Reliability', in Bo Einarsson (ed.), *Accuracy and Reliability in Scientific Computing* (Society for Industrial and Applied Mathematics, 2005).
- 75 Assaad and Williams, 'Technology and Tactics'.
- 76 GG Whitchurch and LL Constantine, 'Systems Theory', in PG Boss, WJ Doherty, R LaRossa, WR Schumm and SK Steinmetz (eds), *Sourcebook of Family Theories and Methods: A Contextual Approach* (New York: Plenum Press, 1993).
- 77 D Saidulu and R Sasikala, 'Machine Learning and Statistical Approaches for Big Data: Issues, Challenges and Research Directions', *International Journal of Applied Engineering Research* 12, no. 21 (2017): 11691–11699.
- 78 AD Larson and CC Hayes, 'An Assessment of Weasel: A Decision Support System to Assist in Military Planning', *Proceedings of the Human Factors and Ergonomics Society* 49, no. 3 (2005): 287–291.
- 79 Harper, 'Centcom Commander Touts Use of AI in Fight against Iran during Operation Epic Fury'.
- 80 Assaad and Williams, 'Technology and Tactics'.
- 81 C Nugent and P Cunningham, 'A Case-Based Explanation System for Black-Box Systems', *Artificial Intelligence Review* 24 (2005): 163–178.
- 82 Zena Assaad and Christine Boshuijzen-van Burken, 'Ethics and Safety of Human-Machine Teaming', *TAS '23: Proceedings of the First International Symposium on Trustworthy Autonomous Systems* (Association for Computing Machinery, 2023), pp. 1–8.
- 83 M Wischniewski, N Krämer and E Müller, 'Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-of-the-Art and Future Directions', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2023), pp. 1–16.
- 84 Alexander M Aroyo et al., 'Overtrusting Robots: Setting a Research Agenda to Mitigate Overtrust in Automation', *Journal of Behavioural Robotics* 12, no. 1 (2021): 423–436.
- 85 Daniel Ullrich, Andreas Butz and Sarah Diefenbach, 'The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction', *Frontiers in Robotics and AI* 8 (2021).
- 86 Dorsey and Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle', p. 187.
- 87 Dan Ariely and Dan Zakay, 'A Timely Account of the Role of Duration in Decision Making', *Acta Psychologica* 108 (2001): 187–207.
- 88 Zena Assaad, 'Artificial Urgency: Reflecting on AI Hype at the 2026 REAIM Summit', *Just Security*, at: www.justsecurity.org/132504/ai-hype-2026-reaim-summit.
- 89 Michael Biesecker, Sam Mednick and Garance Burke, 'As Israel Uses US-made AI Models in War, Concerns Arise About Tech's Role in Who Lives and Who Dies', *AP News*, 18 February 2025.
- 90 This statement was made by Husham Ahmed, Counsellor at the Permanent Mission of Pakistan to the UN in Geneva, during the informal exchanges on AI in the military domain convened by the UN Office for Disarmament Affairs (UNODA) from 15 to 17 June 2026 in Geneva. Formal permission was granted for attributing this statement to Husham Ahmed.
- 91 Yachot, "'Data Is Control'", *The Guardian*, 31 December 2025.
- 92 K Klonowska and TK Woodcock, 'Rhetoric and Regulation: The (Limits of) Human/AI Comparison in Legal Debates on Military AI', in B Boutin, TK Woodcock and S Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (The Hague: MC Asser Press, 2026).
- 93 Hyunwoo Kim, Harin Yu and Hanau Yi, 'The LLM Fallacy: Misattribution in AI-Assisted Cognitive Workflows', *arXiv*, arXiv:2604.14807, at: <https://doi.org/10.48550/arXiv.2604.14807>.

/ PATHWAYS TOWARDS ENSURING HUMAN CONTROL AND JUDGEMENT ACROSS THE MILITARY AI LIFECYCLE: AN AUSTRALIAN ARMY PERSPECTIVE

Ingvild Bode and Anna Nadibaidze

Introduction

Militaries around the world are increasingly integrating various types of AI technologies into weapon systems and forms of military decision-making, including in relation to targeting. AI techniques that may be used for such applications include computer vision systems, natural language processing, large language models (LLMs) and big data analytics.¹ In parallel to this accelerating trajectory, the past decade has seen a diversified governance discussion across several international forums. This discussion is characterised by both formally institutionalised processes at the United Nations (UN), especially the Group of Governmental Experts on Lethal Autonomous Weapon Systems (GGE on LAWS) and more ad hoc, multi-stakeholder initiatives such as the Responsible AI in the Military Domain (REAIM) summits.² The discussion has matured into lists of potential governance principles in the form of voluntary commitments.³ This article focuses on the principle that pertains to maintaining the central role of humans in the use of force.

From the beginning of the international debate, stakeholders have been invested in examining how the use of various AI technologies affects the role of humans in military decision-making. Consequently, the importance of safeguarding the human role is mentioned as a central concern in practically all outcome documents pertaining to autonomous weapon systems and AI in the military domain.⁴ Various terms have been used to describe this human role, with the most prominent being human control—

sometimes with the qualifier ‘meaningful’—and human judgement. The June 2026 version of the GGE on LAWS rolling text, for example, refers to ‘human judgement and control’ in combination.⁵ At the same time, ‘human control and judgement’ remains contested as a term and lacks a universal definition. Many actors, including from Australia, have expressed reservations about employing the specific term ‘human control’.⁶ In this article, we do not take this concept to imply either constant or homogenous types of human supervision in decision-making about the use of force. We understand human control as the capacity to exercise agency: to reasonably understand and foresee an AI system’s functions and effects in a relevant context, to deliberate on context-appropriate actions, and to act on these deliberations.⁷ Focusing on human agency as an integral component of human control and judgement allows us to move away from the binary assumption of a hierarchical relationship of humans and AI technologies.⁸ Broadly understood as ‘the ability to take actions or affect outcomes’,⁹ the term ‘human agency’ captures a nuanced account of the relationship between humans and AI technologies, including how types and degrees of human–AI interaction influence the capacity for any exercise of human agency.¹⁰

Given that safeguarding human control and judgement is such a longstanding concern, there have been some efforts directed at turning this principle into practice.¹¹ Many of these efforts recognise that securing human control and judgement cannot take place at the stage of employing particular applications of AI alone but needs to happen across the whole lifecycle of AI systems, and especially at the early lifecycle stages of pre-development, design and procurement, as well as testing, evaluation, verification and validation (TEVV).¹² Securing the exercise of human control and judgement will therefore involve various groups of humans who exercise their agency in a way that is appropriate to the context.¹³

This paper sketches concrete pathways for implementing the widely recognised principle of human control and judgement in the use of force through highlighting sets of best practices across the AI lifecycle for policymakers, military personnel and developers.¹⁴ These pathways are based on work conducted by the AutoPractices project, which brought together 49 diverse stakeholders to co-create a best practices toolkit to sustain the human role in the context of using AI-based systems in the military domain.¹⁵ To make the potential application of these practices more concrete, we connect their exercise to the case of AI-based decision support systems (AI DSS). We define AI DSS as ‘tools that use AI techniques to collect and analyse data and provide information about the operational environment, as well as actionable recommendations, with the aim of aiding military decision-makers’.¹⁶

The remainder of the paper is structured as follows. First, we highlight why human control and judgement continues to occupy such a central place within debates about AI in the military domain, foregrounding safety, legal and ethical reasoning. Second, we situate the Australian Government’s position regarding human control and judgement in the

context of integrating AI in military decision-making, including Australia's participation in the international governance debate. Third, we review how sets of best practices along the AI lifecycle can contribute to operationalising human control and judgement through unpacking its safety, legal and ethical components. We focus, in individual sub-sections, on best practices that can be performed by Australian policymakers, military personnel and developers within the context of developing and using AI DSS, as well as potential points of tension and implications for the Australian Department of Defence (Australian Defence). Finally, we conclude by sketching open questions and areas of inquiry.

The Importance of Exercising Human Control and Judgement: Safety, Legal and Ethical Purposes

Securing the retention of human control and judgement in military decision-making pertaining to the use of force continues to receive broad support.¹⁷ For example, at the March 2026 meeting of the GGE on LAWS, state delegations characterised the human dimension as central to the group's mandate of proposing 'a set of elements of an instrument ... to address emerging technologies in the area of LAWS'.¹⁸ It is important to underline this enduring importance, because the governance debate around AI in the military domain is under stress at the time of writing (June 2026). The United States, for example, moved away from its previous commitment to soft governance and explicitly suggested removing references to human control and judgement from the GGE on LAWS rolling text at the March 2026 meeting. Simultaneously, the 2026 REAIM Summit outcome document was only endorsed by 39 states, compared to the 61 states who had endorsed the 2024 REAIM Blueprint for Action. The AI governance debate is therefore also affected by a deteriorating global security situation and an ongoing rupture in the international rules-based order, visible in the strains put on the regulatory capacity of international law.

The importance of the human control and judgement principle has become an almost intuitive reference point in the international debate, but it is important to spell out clearly why this is the case.¹⁹ As Davidovic highlights, policies implementing this principle need to be precise—both about what types of human roles there are across the AI system lifecycle and about what purpose the human role serves.²⁰ We unfold safety, legal and ethical purposes of ensuring the exercise of human control and judgement. As described in the introduction, we consider human agency as an integral component of enabling human control and judgement. It is particularly important to ensure that AI DSS are employed in a way that strengthens humans' ability to act and affect outcomes, instead of undermining their potential to exercise their agency.²¹

Safety purposes: Retaining human control and judgement is often characterised as an important safety guardrail in deploying AI technologies in the military domain. This is essentially a functional argument that is tied to the continued concerns around both the reliability and predictability of AI technologies.²² For example, LLMs inevitably produce

factually incorrect information as part of their core statistical design.²³ The purpose of retaining human control and judgement aligns with that of other safety-critical domains where the human being adds value in compensating for LLM unreliability and spotting incorrect information.²⁴

Legal purposes: Across the long-running legal debate, two (somewhat) distinct purposes of human control and judgement have been spelled out. First, some contend that such a principle is already part of the spirit of international law, including international humanitarian law (IHL), albeit in implicit form. As Heyns argues, ‘since the use of force throughout history has been personal, there has never been a need to make this assumption explicit’.²⁵ As the very institution of (international) law and justice is therefore human-centric,²⁶ the role of the human needs to be maintained for the purpose of overall institutional stability.²⁷ Second, others argue that the exercise of human control and judgement is the basis for and a means to ensure compliance with international law,²⁸ as ‘only natural persons are capable of administering the performance of IHL obligations binding on States’.²⁹

Ethical purposes: Centrally retaining the exercise of human control and judgement ensures the continued presence of ethical and moral reasoning. This serves the purpose of countering the risk of a technological systematisation of violence associated with AI technologies in the military domain.³⁰ This line of reasoning underlines that warfare is a human activity that is characterised by vast moral dilemmas and therefore requires contextual human decision-making. Having unfolded the safety, legal, and ethical purposes that human agency serves, we now highlight Australia’s approach to AI in defence and its positioning in the global governance debate.

Responsible AI and Human Control in Australian Defence

Australia is both a key developer of AI in defence and a prominent participant in governance debates on AI in the military domain.³¹ Its 2026 and 2024 National Defence Strategies highlight that ‘as a middle power, Australia must seek military advantage in innovative ways, including through developing asymmetric advantage’,³² which includes developing technologies grouped under the term ‘AI’. This objective is also stated in the 2024 Defence Innovation, Science and Technology Strategy, which describes innovation as key for Australia’s strategic position as a middle power in the Indo-Pacific great power competition.³³ As part of this strategic positioning in the regional and global orders, Australia’s approach seeks to particularly leverage AI for human–machine teaming, as well as to innovate and generate asymmetric advantage.

Many of the AI capabilities that Australian Defence has been developing, testing and demonstrating focus on AI technologies being used as a tool to support military personnel. In January 2026, the Australian Government announced an investment

of almost AU\$40 million in collaborations with research institutions for ‘decision advantage’.³⁴ Official documents and discourses highlight that the main objective is not to replace humans but to assist Australian Army personnel in decision-making, as this is what is perceived to give strategic advantage.³⁵ As highlighted by Wing Commander Kate Conroy during Exercise Talisman Sabre 2025, which involved an AI-based system tested on a flight of the P-8A Poseidon aircraft, ‘The AI adds value when the crew need it, and helps them make faster, smarter decisions ... AI is a helper, not a decider’.³⁶ The Royal Australian Air Force has outlined its vision of becoming ‘AI-ready’ and developing a ‘concept of human-machine teaming’ where AI DSS play a key role.³⁷ One of the systems disclosed by the Air Force is the cognitive assistant avatar Sergeant AIMEE, a chatbot that is portrayed as being useful in the education and training of personnel.³⁸

At the same time, Australia has been actively involved in global debates on potential governance of AI in defence, such as discussions on LAWS held at the UN under the framework of the Convention on Certain Conventional Weapons. Throughout its participation in UN expert meetings on LAWS, Australia has highlighted its commitment to the principle of safeguarding the human role in the use of force, although without supporting new legally binding measures on human control and judgement in the development and use of autonomous weapons. It ‘considers “human control” or “human involvement” as one means—but not the only means—of ensuring compliance with’ international humanitarian law, adding that the ‘level of human control or human involvement will vary depending on the context of the use of a LAWS’.³⁹

Australia has also endorsed all the outcome declarations from the three REAIM summits: the Call to Action (2023), Blueprint for Action (2024) and Pathways to Action (2026). It joined the 2023 US Political Declaration on Responsible Military Use of AI and Autonomy.⁴⁰ Moreover, it voted in favour of UN General Assembly resolutions on LAWS (2023 and 2024) and on ‘Artificial intelligence in the military domain and its implications for international peace and security’ (2025). As noted in the introduction, most of these initiatives include a reference to safeguarding the human role, although these references might be formulated using different terms.⁴¹ For example, the 2026 REAIM Pathways to Action document mentions that ‘AI-enabled decision support systems should support, not replace, the exercise of human judgement’.⁴²

Building on Australia’s commitments, Australian documents have proposed practical ways of implementing the principle of human control and judgement in the use of force, including in the context of autonomous weapon systems. In its 2019 non-paper submission to the GGE on LAWS, Australia outlined the concept of ‘system of control’ applying to ‘all aspects of a weapon system from design through to engagement’, which it also implements in its use-of-force procedures.⁴³ This system is meant to enable ‘controls to be tailored to the specific AWS and its unique operating environment’.⁴⁴ This non-paper notes that Australia finds the specific term ‘human control’ to insufficiently ‘cover the

plethora of practical controls or systems utilised by states and their militaries'.⁴⁵ However, the proposal for the system of control emphasises the need to implement the human role⁴⁶ and to ensure that the use of weapon systems is 'driven by human direction'.⁴⁷ It outlines control mechanisms as part of various stages of the lifecycle of autonomous weapons, ranging from initial political decision-making ('legal and policy framework') to post-weapon use review ('after-action evaluation'). While this framework of control concerns the case of weapon systems, it signals Australian Defence's commitment to engage in practical efforts to operationalise responsible AI in the military domain, as well as to implement a lifecycle-based approach that stretches beyond the stage of employing weapon systems.⁴⁸

The Australian Government also highlights its objective to adopt a comprehensive framework spanning different domains, noting that all technologies, 'including AI, are subject to multi-layered control measures, comprised of technical, legal and doctrinal measures including test and evaluation, and capability-specific training and policies'.⁴⁹ From this perspective, there are various control measures that may be implemented at different stages of the lifecycle, depending on the context of development and intended use.⁵⁰ Australia's Policy Settings for Responsible Use of Artificial Intelligence in Defence, which 'implement Australia's commitments' under the REAIM output documents and the US Political Declaration, state that the responsible use of AI 'must be considered at all stages of the technology life cycle'.⁵¹ This includes, according to these policy settings, risk-based control measures to 'identify, assess and mitigate' risks related to AI, 'at every stage of the technology lifecycle'.⁵² As its delegation noted during the 2026 UN Informal Exchanges on Artificial Intelligence in the Military Domain, Australia supports an approach where responsible use of AI 'must be integrated from the early lifecycle stages'⁵³ and where 'effective oversight throughout the lifecycle is critical'.⁵⁴

In sum, Australia's strategic and diplomatic documents underline that the development and acquisition of AI systems is meant to support military personnel's decision-making. At the same time, its practices in relation to responsible AI and governance debates demonstrate its official commitment to a context-appropriate exercise of human control and judgement across the lifecycle of AI systems. The Australian Government has indicated its interest in moving towards 'practical efforts' to implement responsible development and use of AI in defence.⁵⁵ Getting the right balance between innovation and robust governance throughout the AI lifecycle is therefore a 'key priority for Australia'.⁵⁶ In the next section, this article explores sets of best practices across the AI lifecycle that may be relevant to ongoing efforts within Australian Defence, considering its commitment to a lifecycle-based approach.

Practices Contributing to the Exercise of Human Control and Judgement Across the AI Lifecycle: the Case of AI DSS

Various efforts to operationalise the human control and judgement principle draw on AI lifecycle frameworks applied to the military domain—including in Australia, as demonstrated in the previous section. Lifecycle-based frameworks combine understandings across the engineering lifecycle and the military targeting cycle, while also being attentive to the inherent socio-technical dynamics.⁵⁷ Taken as a whole, a lifecycle framework depicts multiple avenues for exercising human agency and thereby ensuring human control and judgement across the recurring stages of AI development for military contexts, encompassing design, training, testing, procurement, deployment and post-use review. We follow a simplified version of the lifecycle framework developed by the IEEE Standards Association Research Group on Issues of AI and Autonomy in Defense Systems.⁵⁸ This model distinguishes between four broad stages:

- (1) Plan:** this includes defining the scope and determining the objectives of a system
- (2) Design:** this involves outlining specific requirements, architectures, functions, interfaces etc. of the system
- (3) Develop:** this involves the actual building of the system, which includes TEVV
- (4) Deploy and post-use:** this involves integrating the system into operations, which includes maintenance and post-use assessment throughout the lifecycle of the system.⁵⁹

We review different practices applicable to three core sets of stakeholders—policymakers, military personnel, and developers—with a focus on the context of developing and employing AI DSS. In the fourth sub-section, we reflect on the potential tension between some of the practices and the importance that Australian Defence attributes to speed and optimisation.

Policymakers

Policymakers are central focal points for the exercise of human agency as a key dimension of human control and judgement across both ‘stage 1: plan’ and ‘stage 4: deploy’. We highlight three selected practices for policymakers to perform at these stages. These practices not only mark important types of human agency in their own right but also have knock-on effects, like all practices at early AI lifecycle stages, for subsequent exercises of human control and judgement along the lifecycle.⁶⁰ Therefore, such policymaker practices play a central role in ensuring that any employment of AI DSS across a military organisation reflects deliberate, explicit choices made early on, rather than having those determinations pushed to later stages of the lifecycle or even left to happenstance.

Set clear objectives and boundaries for the AI DSS to be developed or acquired (stage 1: plan). Before the development of any AI DSS or other capability, political and strategic decision-makers set as clear a plan as possible around foundational system components. This includes considering at least three aspects: (a) the clear and concrete goals that the AI DSS is meant to serve, (b) planned contexts and environments for the use of the AI DSS, and (c) whether the development and use patterns of the AI DSS align with broader political, strategic, ethical and societal considerations. This step recognises that such foundational parameters of system design amount to socio-political questions that require making explicit, recorded choices.

Setting clear objectives and boundaries for the AI DSS before development and acquisition serves safety, legal and ethical objectives. In terms of safety, it aligns with engineering practice in safety-critical settings, which includes documenting a concept of operations (CONOPS) ‘describing the characteristics and intended usage of a proposed or existing system from the viewpoint of its users’.⁶¹ This is intended to align expectations across different stakeholder groups involved throughout the AI DSS lifecycle and therefore creates an important throughline of communication and expectation.

Going through this practice as a process should also serve to create human-set guardrails around legally and ethically defined undesirable ways of employing AI DSS. In legal terms, any AI DSS that ‘contribute[s] to the conduct of hostilities (e.g., target identification, collateral damage estimation, selection of weapons) fall[s] within the scope of Additional Protocol I (AP I) Article 36 legal reviews’.⁶² This means that the central question of whether the use of AI DSS may, even inadvertently, lead to indiscriminate targeting, which is clearly prohibited under IHL, needs to be raised and addressed at this early stage.⁶³ In ethical terms, this practice could serve to set limits on the use of AI DSS that exceed positive legal obligations from the get-go, such as clear commitments to higher levels of civilian protection or exercising a higher degree of precautions as a result of using particular AI DSS. Highlighting these vital legal and ethical considerations should therefore also entail identifying use cases and applications where AI DSS are unsuitable. This has the potential to counterbalance the detrimental kneejerk response that AI DSS are the solution for every possible identified capability need.

Design a comprehensive risk assessment framework and make (political) choices regarding risk acceptance and risk thresholds (stage 1: plan). This practice involves constructing a framework to evaluate various risks associated with employing a particular type of AI DSS and using such a framework as a guide throughout the AI DSS’s lifecycle. Assessing risks associated with AI DSS is not only an important exercise of human agency by policymakers before such a system is even developed. It also ties policymakers to setting clear benchmarks to ensure that AI DSS are used for intended objectives and enable retaining human control and judgement.

There is currently no risk assessment framework that is specific to AI DSS in the military context, at least not in the public domain. Such a risk assessment framework could be designed to incorporate quantitative and qualitative evaluations of risk,⁶⁴ organised into tiers that classify AI DSS according to their risk profile—for example high-risk, medium-risk and low-risk categories.⁶⁵ Particular types of AI DSS would be classified into these risk categories depending on factors such as intended purpose of use (e.g., targeting, command-and-control, logistics), intended context of use (e.g., populated/unpopulated areas) or data quality. These factors are key in delineating high-risk, medium-risk and low-risk categories because they matter in assessing the risk of physical harm as well as the risk posed to civilian populations that can be associated with using particular types of AI DSS.

As risk assessment frameworks inform the practices of actors across the entire lifecycle, we consider them to be a particularly vital way of exercising human control and judgement. Thinking about AI DSS in terms of comprehensive risk assessment frameworks and the human roles therein serves legal, ethical and safety purposes. It follows that classifying a system as low, medium or high risk triggers different measures for the design, deployment and use of such systems. To illustrate, we focus on a high-risk AI DSS where the stakes are highest.

At the point of planning the design of high-risk AI DSS, policymakers need to consider how to minimise civilian harm. Such planning should include design provisions for the AI DSS—for example, the capacity to recognise and track civilian populations and alert forces to their presence.⁶⁶ Such provisions could also elevate ethical responsibility in using AI DSS,⁶⁷ thereby explicitly integrating the potential humanitarian benefits of such systems in proactive ways. This could entail, for example, prioritising civilian protection as a positive duty and featuring procedures for human control along the lifecycle in ways that better address unintended but foreseeable harms. At a more fundamental level, such measures should also include working towards actively addressing the higher likelihood of enemy dehumanisation and countering the risk of moral numbness in the operator that has been analytically associated⁶⁸ and empirically reported in the use of AI DSS.⁶⁹

Systems with high risk profiles would necessitate greater human intervention.⁷⁰ This could involve, for example, rigorous scrutiny and careful review processes where human operators of AI DSS cross-verify the system's outputs before acting on them. This serves the legal purpose of ensuring that the use of a particular AI DSS in a particular context complies with core IHL principles, particularly the principle of precaution. A high risk profile should also involve slowing down the overall decision-making speed. AI technologies can allow military personnel to access a much wider array of sources in making crucial decisions—for example, legal judgements—around key principles such as precautions and proportionality. This also echoes longstanding military practices around tactical patience, clearly acknowledging that speeding up decision-making across the board cannot be an end in itself.⁷¹

Ensure that capability needs and objectives are met by the products offered

(stage 4: deploy). This practice is at the heart of procurement processes, serving to ensure that critical choices made by policymakers at stage 1 are implemented at the moment of acquiring a specific AI DSS, rather than being displaced by shifting vendor priorities.⁷² While maintaining such a throughline of policymaker decisions from planning to procurement will generally be central, the risk of the widespread hype narratives around AI technologies shaping procurement decisions warrants particular attention.⁷³ An important element of this practice would be to engage in open conversations with diverse potential vendors rather than becoming locked into working with a limited range of suppliers circulating (potentially) hyped products. Unpacking the hype narratives should also translate into careful communication of TEVV results to military commanders and operators who will use the procured AI DSS.

At its heart, this practice therefore concerns safety requirements by ensuring that relevant military personnel have access to essential information needed to comprehend how specific AI DSS operate within their intended operational contexts. Yet the ability to pierce hype narratives around AI technologies associated with this practice also has important ethical implications because it can contribute to maintaining a sense of moral agency among military personnel using AI DSS. It can allow military personnel to retain a sense of the moral trade-offs that are inherent in warfare and avoid falling into the techno-solutionist trap that ‘armed conflict can be remade into manageable problems and “solved” once the right algorithm is developed’.⁷⁴

Military Personnel

Military personnel are the central actors across ‘stage 4: deploy and post-use’ and their practices have long been a focal point for sustaining human agency.

Develop comprehensive training programs and protocols for all military personnel that either employ or recommend employing AI DSS. Educating humans, including but not limited to operators and commanders, about relevant AI systems and how they are integrated into a network (which is often the case for AI DSS) plays a key role in retaining adequate levels of human control. Training should be comprehensively structured. It should include technical training around, for example, AI literacy—in other words, a basic technical and conceptual understanding of AI systems.⁷⁵ Such basic training should be seen as a premise for ensuring that military personnel can fulfil both legal and safety functions in exercising their role. Yet training should also cover human–machine interaction across socio-technical components and how humans engage with AI systems. This type of training holds particular significance as it helps cultivate and deepen awareness and understanding among military personnel of decisive phenomena including automation bias, cognitive biases and sycophantic behaviour.⁷⁶

To counteract inherent biases and assumptions embedded in human–machine interaction, a sustained training program addressing both the dynamics of such interactions and their effects on human control is essential.⁷⁷ This type of training should therefore provide broader, more contextualised knowledge about how AI DSS work in practice and how this shapes the decision-making space for military personnel. This knowledge is likely to be particularly crucial for legal and ethical purposes as it centres on the retention and sustenance of human agency. This requires training that takes a deeper dive into the underlying, socio-technical logics of AI DSS.

Adequate comprehension of AI DSS and how humans interact with them should strike a balance that guards against users becoming overly confident or reliant on such technologies. Continuous training initiatives also aim to inform involved military personnel about the potential ramifications of deploying AI DSS in different contexts, spanning strategic, legal and ethical implications. Such knowledge enhances humans' capacity to intervene effectively and reflect on, dispute or reject AI DSS recommendations as necessary and appropriate to the context—all qualities that underlie legal and ethical purposes that human control and judgement need to fulfil. A comprehensive training regime, moreover, should include recognition that its delivery and curriculum will have to be adaptive and evolve alongside any technological and operational changes.

Establish ways for military personnel to understand when AI DSS do not work according to planned objectives and requirements. The history of integrating AI predecessor technologies (such as automation and autonomy) exhibits various examples where known failure modes were not (adequately) communicated to the human operators of such systems.⁷⁸ Yet this is an essential component that allows humans to decide how best to act in cases of system malfunctions, through training protocols around steps to follow in unexpected situations. This practice also involves alerting military operators to blind spots in the AI DSS training data, as well as edge cases. This practice can be primarily associated with the safety purposes of human control, as it can enable military personnel to prevent mistakes from happening.⁷⁹

It can therefore be an important step towards strengthening human operators' ability to understand and assess the outputs of AI DSS. It also provides an important counterbalance against potential risks of human de-skilling and cognitive dependence that have been found in research around human–AI interaction in the civilian domain.⁸⁰ Such a practice also encompasses generating awareness around potential effects of machine–machine interaction—especially when various AI DSS may be used as part of a complex system of various intersecting types.⁸¹

Institute feedback loops between military personnel and developers across the AI DSS lifecycle. This serves to sustain the exercise of human control and judgement in providing a consistent throughline of information from end user to developer. It ensures

that key design choices fulfil their intended objectives and enable rather than delimit human agency. This practice is based on the ability to discuss particular uses of AI DSS in context and on military personnel sharing experiences of interacting with AI DSS. Creating such a throughline of communication serves primarily both safety and legality purposes. In safety terms, it allows reporting of potential malfunctions and making assessments about their relative significance in an operational context. In legal terms, it provides important information about whether design choices actually had the intended effect in terms of sustaining human control and judgement. The impact of this practice ultimately rests on militaries maintaining an organisational culture that is somehow open to feedback and exchange.

Developers

Practices contributing to the exercise of human control and judgement are applicable to developers, including those from the private sector and research institutes collaborating with Australian Defence.⁸² As the Global Commission on Responsible AI in the Military Domain notes, industry actors' 'decisions significantly influence whether downstream users can integrate [AI] technologies in ways that align with ethical, legal, and operational standards'.⁸³ The role of developers of AI systems such as AI DSS is particularly central to stages 2 (design) and 3 (develop).

Incorporate feedback from users via communication mechanisms. As mentioned in the last practice listed in the previous section, appropriate communication mechanisms between end users and developers are crucial to ensure that AI DSS are developed to match the operational and strategic objectives, as well as meeting legal obligations and ethical commitments. Feedback from military personnel is useful in making design choices about AI DSS and developing the systems, as ultimately the systems are meant to enhance the users' capability to exercise their agency according to the appropriate context. This includes involving users in the design of the interface, particularly to ensure that these design choices match the intended role of AI DSS and their operational value within the context. This practice fulfils primarily a safety purpose, as its main objective is securing alignment between the development of AI DSS and their intended uses and addressing risks associated with those contexts. At the same time, institutionalising such feedback mechanisms also contributes to enhancing military personnel's ability to engage in legal decision-making in their use of AI DSS, particularly in relation to IHL principles which require the exercise of contextualised human control and judgement.⁸⁴

At stage 2, design AI DSS interfaces that present information in a manner appropriate to the context, while considering the limitations and risk assessment frameworks set previously (at stage 1). The way in which interfaces present information to human decision-makers deserves particular attention, as interfaces are meant to strengthen users' capacity to consider various options, deliberate on courses of action,

and decide on the one(s) most appropriate to the context. Developers should pay attention to interface design because it has potential to enable and constrain the exercise of human agency and contribute to safety, legal and ethical considerations.⁸⁵

For example, interfaces that constantly change the way they present information or that prime users towards a particular course of action would be less likely to enable users' agency at later stages, including at stage 4. Interface designs that feature a balance between necessary information and cognitive overload contribute to military personnel's ability to exercise control and judgement from safety, legal and ethical standpoints. A balanced approach to interface design that includes some level of preselection but does not lean into oversimplification is more likely to empower users of the AI DSS. Design choices at stage 2 should incorporate decisions made at stage 1, including the clear objectives and boundaries of AI DSS and any risk assessment frameworks identified by relevant policymakers. Interfaces of AI DSS should also incorporate ways for users to be aware of when systems are not performing as expected.

Conduct TEVV procedures that incorporate concrete metrics on various dimensions (stage 3). TEVV procedures should incorporate concrete metrics on a variety of aspects, especially on whether intended users are empowered to engage in decision-making that is compliant with IHL principles and appropriate ethical frameworks, as well as operational and security objectives. This practice therefore fulfils safety, legal and ethical purposes associated with the exercise of human agency, which contributes to human control and judgement. Testing of AI DSS and their components should ultimately ensure that intended users possess the appropriate mechanisms to deliberate on actions in relation to not only expected conditions of use but also edge cases or unexpected scenarios. Comprehensive procedures that match the limitations and risk assessment frameworks set previously (at stage 1) are particularly crucial in TEVV. Australia's system of control framework from 2019 also highlights the importance of comprehensive testing and evaluation in autonomous weapon systems.⁸⁶

At stage 3, test the AI DSS interfaces to identify and document limitations.

Developers can enable intended users' capacity to exercise agency by involving these users in the testing of the interface and monitoring what the use of these AI DSS means for military personnel's decision-making in a particular context. In other words, it is beneficial for testing practices to include aspects of human-machine interaction, including dynamics such as automation biases or sycophancy, rather than exclusively focusing on the technical performance of AI DSS. This practice builds on other interface design related practices mentioned earlier, underlining the importance of consistency and alignment across the lifecycle stages.⁸⁷

Ensure that appropriate datasets are used and document potential limitations or biases in data. Across both stages 2 and 3, developers can engage in many practices that ensure transparency surrounding the data used to train the AI DSS. These practices are important as they strengthen users' ability to understand how AI DSS may (mal) function and reasonably foresee the impact of using these AI DSS within particular contexts. Developers should take measures to improve data quality as much as possible and record limitations, as part of both design and testing. Documentation regarding training data is a way to provide transparency surrounding potential biases in the data that the AI DSS were trained on. Such documentation informs other humans across the lifecycle, such as military personnel and policymakers, about limitations to the data and strengthens their ability to judge whether AI DSS are appropriate to a certain context. There are several key questions regarding the training data, including concerning the selection, labelling and assessment of the data involved in the development of the AI DSS. This practice is particularly relevant for the safety purposes of ensuring the exercise of human control and judgement. It seeks to mitigate safety risks involved with incomplete or wrong data in a military context, which may exacerbate security risks such as the brittleness of AI DSS.⁸⁸

Implications for Australian Defence

The practices outlined in the previous section contribute to operationalising the principle of human control and judgement across the lifecycle of AI DSS used in the military domain. They are therefore valuable to Australian Defence as it seeks to both promote innovation in its armed forces and implement its commitment to ensuring the human role in the use of force. Many of these practices also match the approach adopted in Australia's 'system of control' 2019 non-paper, such as comprehensive TEVV, training of personnel, setting specific parameters of use, and conducting post-use reviews.⁸⁹ Although Australia's 2019 proposal is specific to the case of weapon systems, and the practices listed here focus on the development and use of AI DSS, there are many points of intersection that are relevant to Australian Defence.

At the same time, the application of some of these practices may entail points of discussion, tension and even resistance. In certain contexts, a comprehensive risk assessment framework which would involve human intervention, review processes and cross-checking mechanisms can be considered an obstacle to reaping the benefits of employing AI DSS. The development and the use of AI DSS are associated with speed, scale and optimisation.⁹⁰ Mechanisms that would slow down the tempo of decision-making are associated with reduced efficiency, which would be seen as countering the priority attributed to developing AI technologies in defence. As Australia's 2026 National Defence Strategy states, 'Defence must keep pace with these changes [in emerging technologies including AI] to leverage opportunities, respond to evolving threats and

remain interoperable with allies and partners'.⁹¹ There is therefore a potential point of tension between operationalising the role of the human on one hand, and the strategic importance given to increasing the armed forces' efficiency and the speed of decision-making on the other.

To meet Australian Defence's strategic objectives of developing asymmetric advantage and bolstering 'the ADF [Australian Defence Force]'s ability to deter and respond',⁹² the use of AI DSS should 'extend users' will' rather than undermine it.⁹³ This extension of the will—and the use of AI to truly support and help operators—requires humans across the lifecycle to exercise their control and judgement in a way that matches Australia's priorities as well as increasing its interoperability with allies.⁹⁴ Ultimately, the exercise of human agency which contributes to human control and judgement in the development and use of AI DSS involves a set of common practices but not a 'one-size-fits-all model'.⁹⁵ Each set of actors, whether Australian policymakers, military personnel or developers/industry, will therefore have to navigate respective trade-offs to ensure the exercise of human control and judgement.

Conclusion

Global governance debates on military applications of AI are considering the operationalisation of high-level principles agreed across various forums. As a key developer of AI in defence and an active participant of governance debates, Australia has demonstrated its commitment to the principle of the human role in the use of military force and to pursuing the debate on how to implement it in practice. This article has presented security, ethical and legal reasons why this work should continue and has offered sets of practices applicable to various groups of humans—policymakers, military personnel, and developers—within the context of development and use of AI DSS. These practices are based on the findings of the AutoPractices project, which co-created a toolkit of best practices to strengthen the exercise of human control and judgement across the lifecycle of military AI systems.

The practices included in this overview match the Australian approach to the human role presented in its 2019 non-paper, even though this proposal states that the term 'human control' is not the most useful. As Trabucco notes, Australia's approach broadens the system's timeline and 'lengthen[s] the perspective horizontally to broaden the range of actors and decision-making capacities and fold[s] them into a network of human control'.⁹⁶ The practices explored in this article can also be integrated as part of a networked approach to the exercise of control and judgement across human actors involved at different stages of the lifecycle, with a particular focus on the early stages (1–3). While there are potential points of tension or even resistance to some practices,

this article argues that strengthening this network of practices would assist Australia in achieving its objective of employing AI to gain a decision advantage while remaining committed to legal obligations and safety considerations, especially as achieving this balance is a 'key priority' for Australia.⁹⁷

Acknowledgements: We are grateful for the constructive feedback we received on earlier drafts of the article from Nicholas Carroll and Neil Renic at the Special Issue workshop, as well as from Vincent Boulanin as part of the review process. We also thank all the stakeholders who participated in the AutoPractices project process.

Funding acknowledgements: This research was supported through funding by the European Research Council via the European Union's Horizon 2020 research and innovation program (grant number 852123, the AutoNorms project; and grant number 101156237, the AutoPractices project) and the Independent Research Fund Denmark (grant ID 3119-00023B, the HuMach project).

About the Authors

Dr Ingvild Bode is a Professor of International Relations at the University of Southern Denmark (SDU). At SDU, she is also Director of the Center for War Studies (CWS). She is currently a fellow with the UN Institute for Disarmament Research (UNIDIR) and an associate fellow with the Royal United Services Institute (RUSI). She served as Co-Chair of the IEEE-SA Research Group on Issues of AI And Autonomy in Defence Systems and was formerly an expert on the Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM). Currently, Ingvild is an academic ambassador of the Responsible by Design Initiative led by Article 36 Legal. Her research focuses on processes of policy and normative change, especially with regard to the use of force. Her work is increasingly dedicated to critically examining the roles of technologies in international relations and international security, especially Artificial Intelligence (AI) technologies in and beyond the military domain.

Dr Anna Nadibaidze is a postdoctoral researcher in international politics at the Center for War Studies, University of Southern Denmark (SDU). Her research is part of the Independent Research Fund Denmark-funded HuMach project, as well as the European Research Council-funded AutoNorms and AutoPractices projects. She investigates military applications of artificial intelligence and their implications for international relations and global security. Anna holds a PhD in Political Science from SDU. Her work has been published in journals such as *Contemporary Security Policy*, *Big Data & Society*, *Ethics and Information Technology*, and *European Journal of International Security*.

ENDNOTES

- 1 Nathan G Wood et al., 'Stop Saying "AI"', arXiv:2602.17729, arXiv (2026): 28, at: <https://doi.org/10.48550/arXiv.2602.17729>.
- 2 While the GGE on LAWS debate involves the 128 High Contracting Parties of the UN Convention on Certain Conventional Weapons, the three REAIM summits that have been held to date have been more limited in geographical scope with, respectively, 60, 61 and 39 states endorsing the REAIM 2023, REAIM 2024 and REAIM 2026 summit outcomes.
- 3 Government of the Netherlands, *REAIM 2023 Call to Action*, 16 February 2023, at: www.government.nl/documents/2023/02/16/ream-2023-call-to-action; Ministry of Foreign Affairs, Republic of Korea, *Blueprint for Action*, 11 September 2024, at: www.mofa.go.kr/www/brd/m_4080/down.do?brd_id=235&seq=375378&data_tp=A&file_seq=9; Ministry of Foreign Affairs, Spain, REAIM Pathways to Action, 2026, at: www.exteriores.gob.es/en/REAIM2026/Documents/REAIM%202026%20Pathways%20to%20Action.pdf; UNODA, *GGE on LAWS Rolling Text, Status Date: 05 June 2026* (2026), at: [https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_\(2026\)/GGE_LAWS_-_Rolling_Text_-_5_June_2026.pdf](https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/GGE_LAWS_-_Rolling_Text_-_5_June_2026.pdf).
- 4 Ingvild Bode, *Emerging Norms Around Military Applications of AI: The Case of Human Control*, GC REAIM Expert Policy Note Series (The Hague Centre for Strategic Studies, 2025), at: <https://hcsc.nl/wp-content/uploads/2025/05/Bode-1.pdf>.
- 5 UNODA, *GGE on LAWS Rolling Text, Status Date: 05 June 2026*, para. 12.
- 6 Australia, *Australia's System of Control and Applications for Autonomous Weapon Systems*, submission to United Nations Convention on Certain Conventional Weapons, Geneva, 2019 (UNODA, 2019), para. 42, at: [https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_\(2019\)/CCWGGGE.12019WP.2Rev.1.pdf](https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2019)/CCWGGGE.12019WP.2Rev.1.pdf).
- 7 The AutoPractices Project, *Strengthening Human Agency in the Military Domain: Best Practices Toolkit for Policymakers, Developers, and Users of AI Systems* (Center for War Studies, University of Southern Denmark, 2026), p 6, at: <https://doi.org/10.5281/zenodo.17671715>.
- 8 Bérénice Boutin and Taylor Woodcock, 'Aspects of Realizing (Meaningful) Human Control: A Legal Perspective', in Robin Geiß and Henning Lahmann (eds), *Research Handbook on Warfare and Artificial Intelligence* (Edward Elgar Publishing, 2024); Klaudia Klonowska and Taylor Kate Woodcock, 'Rhetoric and Regulation: The (Limits of) Human/AI Comparison in Legal Debates on Military AI', in Bérénice Boutin et al. (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (T.M.C. Asser Press, 2026); Anna Nadibaidze, 'Investigating the Human Role in Warfare: Beyond "Control" and Towards "Agency"?' *The HuMach Blog*, 17 February 2025, at: <https://humach.eu/investigating-the-human-role-in-warfare-beyond-control-and-towards-agency>.
- 9 Alan Chan et al., 'Harms from Increasingly Agentic Algorithmic Systems', *2023 ACM Conference on Fairness, Accountability, and Transparency*, 12 June 2023, 651–66, at: <https://doi.org/10.1145/3593013.3594033>. We use this brief definition as a pragmatic starting point for agency in the context of this practically oriented paper. The notion of agency has a rich conceptual heritage across various social sciences disciplines that we cannot capture in this contribution. See Werner Rammert, *Distributed Agency & Digital Technology: A Social Pragmatist View on Human-Technology Interaction* (Springer Fachmedien Wiesbaden, 2025).
- 10 Ingvild Bode et al., 'Ensuring the Exercise of Human Agency in AI-Based Military Systems: Concerns Across the Lifecycle', *Ethics and Information Technology* 27, no. 50 (2025), at: <https://doi.org/10.1007/s10676-025-09861-2>.
- 11 Vincent Boulanin et al., *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (Stockholm International Peace Research Institute and International Committee of the Red Cross, 2020), at www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf; Jovana Davidovic, 'Rethinking Human Roles in AI Warfare', *Nature Machine Intelligence* 7, no. 10 (2025): 1593–95, at: <https://doi.org/10.1038/s42256-025-01123-6>; Merel Ekelhof and Giacomo Persi Paoli, *The Human Element in Decisions About the Use of Force* (United Nations Institute for Disarmament Research, 2020), at: https://unidir.org/wp-content/uploads/2023/05/UNIDIR_Iceberg_SinglePages_web.pdf.

- 12 IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (IEEE SA, 2024), at: <https://ieeexplore.ieee.org/document/10707139>.
- 13 Vincent Boulanin and Dustin A Lewis, 'Responsible Reliance Concerning Development and Use of AI in the Military Domain', *Ethics and Information Technology* 25, no. 8 (2023): 1–5, at: <https://doi.org/10.1007/s10676-023-09691-0>; Boutin and Woodcock, 'Aspects of Realizing (Meaningful) Human Control'; Lena Trabucco, 'AI-Enabled Autonomous Weapons and Human Control. Part I: Human Control and Machine Learning Design and Development', *International Law Studies* 106 (2025): 534–78.
- 14 IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, *A Framework for Human Decision-Making Through the Lifecycle of Autonomous and Intelligent Systems*.
- 15 The AutoPractices Project, *Strengthening Human Agency in the Military Domain*; The AutoPractices Project, *Map of Practices* (Center for War Studies, 2025), at: www.autonorms.eu/governing-ai-technologies-in-the-military-domain-from-the-bottom-up-map-of-practices.
- 16 Marta Bo et al., 'Views of Members of the Scientific Community and Civil Society Pursuant to Resolution 79/239 "Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security" Adopted by the General Assembly on 24 December 2024, in Accordance with the Request of the UN Secretary-General Contained in Note Verbale ODA/2025-00029/AIMD', 11 April 2025, at: [https://docs-library.unoda.org/General_Assembly_First_Committee_-Eightieth_session_\(2025\)/79-239-Dorsey-Bo-Bode-Schwarz-EN.pdf](https://docs-library.unoda.org/General_Assembly_First_Committee_-Eightieth_session_(2025)/79-239-Dorsey-Bo-Bode-Schwarz-EN.pdf).
- 17 Laura Bruun, *Towards a Two-Tiered Approach to Regulation of Autonomous Weapon Systems: Identifying Pathways and Possible Elements* (Stockholm International Peace Research Institute, 2024), at: <https://doi.org/10.55163/LPED7967>.
- 18 UN Convention on Certain Conventional Weapons, *Final Report. Meeting of the High Contracting Parties to the Convention in Certain Conventional Weapons. CCW/MSP/2025/8*, 27 November 2025, at: [https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Meeting_of_High_Contracting_Parties_\(2025\)/CCW-MSP-2025-8_English.pdf](https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Meeting_of_High_Contracting_Parties_(2025)/CCW-MSP-2025-8_English.pdf).
- 19 Rebecca Crootof et al., 'Humans in the Loop', *Vanderbilt Law Review* 76, no. 2 (2023): 429–510.
- 20 Davidovic, 'Rethinking Human Roles in AI Warfare'; Jovana Davidovic, 'On the Purpose of Meaningful Human Control of AI', *Frontiers in Big Data* 5 (2023): 1017677, at: <https://doi.org/10.3389/fdata.2022.1017677>.
- 21 Anna Nadibaidze, 'Do AI Decision Support Systems "Support" Humans in Military Decision-Making on the Use of Force?', *Opinio Juris*, 29 November 2024, at: <https://opiniojuris.org/2024/11/29/do-ai-decision-support-systems-support-humans-in-military-decision-making-on-the-use-of-force>; Alexander Blanchard and Laura Bruun, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting: A Comparison and Recommended Policy Responses* (Stockholm International Peace Research Institute, 2025), at: <https://doi.org/10.55163/YQBY3151>; Nathan Wood, 'Autonomous and AI-Enabled Systems: Extensions or Replacements of Human Will and Control?', *Ethics and Information Technology* 28 (2026): 1–16, at: <https://doi.org/10.1007/s10676-025-09876-9>.
- 22 Brendan Walker-Munro and Zena Assaad, 'The Guilty (Silicon) Mind: Blameworthiness and Liability in Human-Machine Teaming', *Cambridge Law Review* 8, no. 1 (2023); Mary Cummings, 'Prohibiting Generative AI in Any Form of Weapon Control', paper presented at 39th Conference on Neural Information Processing Systems, 2025, at: <https://openreview.net/pdf?id=uEY7kQsiZz>.
- 23 Arthur Holland Michel, *Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making* (International Committee of the Red Cross, 2024), at: <https://shop.icrc.org/decisions-decisions-computation-and-artificial-intelligence-in-military-decision-making-pdf-en.html>; Jonathan Kwik, 'Digital Yes-Men: How to Deal with Sycophantic Military AI?', *Global Policy* 16, no. 3 (2025): 467–473, at: <https://doi.org/10.1111/1758-5899.70042>.
- 24 Qiaochu Zhang, 'Navigating the Complexities of Exercising Human Agency in Human-Machine Interaction Across the AI Lifecycle', The AutoNorms Blog, 21 October 2024, www.autonorms.eu/navigating-the-complexities-of-exercising-human-agency-in-human-machine-interaction-across-the-ai-lifecycle.
- 25 Christof Heyns, 'Autonomous Weapons Systems: Living a Dignified Life and Dying a Dignified Death', in *Autonomous Weapons Systems: Law, Ethics, Policy*, ed. Nehal Bhuta et al. (Cambridge University Press, 2016), 8.

- 26 International Committee of the Red Cross, *Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain* (April 2025), www.icrc.org/sites/default/files/2025-04/ICRC_Report_Submission_to_UNSG_on_AI_in_military_domain.pdf; Joanna J. Bryson et al., 'Of, for, and by the People: The Legal Lacuna of Synthetic Persons', *Artificial Intelligence and Law* 25, no. 3 (2017): 273–91; Anna Nadibaidze and Ingvild Bode, 'Five Questions We Often Get Asked About AI in Weapon Systems and Our Answers', *The AutoNorms Blog*, 25 September 2023, at: www.autonorms.eu/five-questions-we-often-get-asked-about-ai-in-weapon-systems.
- 27 Davidovic, 'Rethinking Human Roles in AI Warfare'.
- 28 Arthur Holland Michel, *Autonomous Refusal in Lethal Weapons* (Center for War Studies, University of Southern Denmark, 2026), p. 5, at: <https://doi.org/10.5281/zenodo.17811999>.
- 29 Dustin A Lewis and Hannah Sweeney, *Exercising Cognitive Agency: A Legal Framework Concerning Natural and Artificial Intelligence in Armed Conflict* (Harvard Law School, 2025), p. 10, at: <https://pilac.law.harvard.edu/exercising-cognitive-agency>.
- 30 Neil Renic and Elke Schwarz, 'Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing', *Ethics & International Affairs* 37, no. 3 (2023): 321–343; Atay Kozlovski, 'Meaningful Human Control over AI Military Decision Support Systems: Exploring Key Challenges', in Bernhard Koch and David Winkler (eds), *Artificial Intelligence Ethics in Military Medicine and Humanitarian Healthcare* (Springer Nature Switzerland, 2026).
- 31 Kate S Devitt and Damian Copeland, 'Australia's Approach to AI Governance in Security and Defence', in Michael Raska and Richard A Bitzinger (eds), *The AI Wave in Defence Innovation* (Routledge, 2023); Peter Layton, *Evolution Not Revolution: Australia's Defence AI Pathway*, DAIO Study 22/02 (Defense AI Observatory, 2022), at: https://defenseai.eu/wp-content/uploads/2023/01/DAIO_Study2202.pdf.
- 32 Department of Defence, *2026 National Defence Strategy* (Australian Government, 2026), at: www.defence.gov.au/about/strategic-planning/2026-national-defence-strategy-2026-integrated-investment-program; Department of Defence, *2026 Integrated Investment Program* (Australian Government, 2026), at: www.defence.gov.au/about/strategic-planning/2026-national-defence-strategy-2026-integrated-investment-program, p. 80. See also Department of Defence, *2024 National Defence Strategy* (Australian Government, 2024), at: www.defence.gov.au/about/strategic-planning/2024-national-defence-strategy-2024-integrated-investment-program; Department of Defence, *2024 Integrated Investment Program* (Australian Government, 2024), p. 64, at: www.defence.gov.au/about/strategic-planning/2024-national-defence-strategy-2024-integrated-investment-program.
- 33 Department of Defence, *Accelerating Asymmetric Advantage: Delivering More, Together* (Australian Government, 2024), at: www.defence.gov.au/about/strategic-planning/accelerating-asymmetric-advantage-delivering-more-together.
- 34 Department of Defence, 'Harnessing Tech to Drive Faster Decisions Across Defence', *Defence* (website), 5 January 2026, at: www.defence.gov.au/news-events/releases/2026-01-05/harnessing-tech-drive-faster-decisions-across-defence.
- 35 'Artificial Intelligence to Improve Decision-Making', *Department of Defence* (website), 18 November 2021, at: www.defence.gov.au/news-events/news/2021-11-18/artificial-intelligence-improve-decision-making; 'Enhancing AI to support the warfighter', *Department of Defence* (website), 9 August 2024, at: www.defence.gov.au/news-events/news/2024-08-09/enhancing-ai-support-warfighter.
- 36 'AI Takes Flight on Talisman Sabre', *Department of Defence* (website), 11 August 2025, at: www.defence.gov.au/news-events/news/2025-08-11/ai-takes-flight-talisman-sabre.
- 37 Alexander Manning, 'Getting Ready for Artificial Intelligence', *Department of Defence* (website), 23 November 2021, at: www.defence.gov.au/news-events/news/2021-11-23/getting-ready-artificial-intelligence; Royal Australian Air Force, 'Getting Ready for Artificial Intelligence' (video), *YouTube*, 24 November 2021, at: www.youtube.com/watch?v=J4K71RPcv5U.
- 38 Royal Australian Air Force, 'Getting Ready for Artificial Intelligence'.
- 39 Australian Government, *Australia's Submission to the United Nations Secretary-General's Report on Lethal Autonomous Weapons Systems* (UNODA, May 2024), para. 26, at: [https://docs-library.unoda.org/General_Assembly_First_Committee_-_Seventy-Ninth_session_\(2024\)/78-241-Australia-EN.pdf](https://docs-library.unoda.org/General_Assembly_First_Committee_-_Seventy-Ninth_session_(2024)/78-241-Australia-EN.pdf).

- 40 Defence Ministers, 'Australia Joins Declaration on Safe and Responsible Artificial Intelligence in the Military', media release, 3 November 2023, at: www.minister.defence.gov.au/media-releases/2023-11-03/australia-joins-declaration-safe-responsible-artificial-intelligence-military.
- 41 Bode, *Emerging Norms around Military Applications of AI*.
- 42 Ministry of Foreign Affairs, Spain, *REAIM Pathways to Action*, para. 12.
- 43 Australia, *Australia's System of Control and Applications for Autonomous Weapon Systems*, para. 3.
- 44 Ibid., para. 1.
- 45 Ibid., para. 42.
- 46 Trabucco, 'AI-Enabled Autonomous Weapons and Human Control. Part I', p. 546.
- 47 Australia, *Australia's System of Control and Applications for Autonomous Weapon Systems*, para. 8.
- 48 Department of Foreign Affairs and Trade, *Australia's Submission to the UN Secretary-General's Report on "Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security"* (Australian Government, 7 August 2025), para. 8, at: [https://docs-library.unoda.org/General Assembly First Committee -Eightieth session \(2025\)/79-239-Australia-en.pdf](https://docs-library.unoda.org/General_Assembly_First_Committee_-_Eightieth_session_(2025)/79-239-Australia-en.pdf).
- 49 Ibid., para. 8.
- 50 Lauren Sanders, 'Symposium on Military AI and the Law of Armed Conflict: Bridging the Legal Gap Between Principles and Standards in Military AI—Assessing Australia's "System of Control" Approach', *Opinio Juris*, 3 April 2024, at: <https://opiniojuris.org/2024/04/03/symposium-on-military-ai-and-the-law-of-armed-conflict-bridging-the-legal-gap-between-principles-and-standards-in-military-ai-assessing-australias-system-of-control-appro>.
- 51 Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence* (Australian Government, 2026), pp. 5–6, at: www.defence.gov.au/sites/default/files/2026-03/Policy-Settings-for-Responsible-Use-of-Artificial-Intelligence-in-Defence-%5B%5D.pdf.
- 52 Ibid., p. 9.
- 53 Australian Delegation at the Informal Exchanges on AI in the Military Domain and Its Implications for International Peace and Security, United Nations in Geneva, 15 June 2026. Quote recorded and transcribed by the authors.
- 54 Laura Varella, 'Report on the Informal Exchange on Artificial Intelligence in the Military Domain', in *AWS Diplomacy Report* (Reaching Critical Will, 2026), p. 14, <https://reachingcriticalwill.org/images/documents/Disarmament-fora/aws/2026/reports/AWSR3.1.pdf>.
- 55 Department of Foreign Affairs and Trade, *Australia's Submission to the UN Secretary-General's Report on "Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security"*, para. 8.
- 56 Australian Delegation at the Informal Exchanges on AI in the Military Domain and Its Implications for International Peace and Security, United Nations in Geneva, 15 June 2026. Quote recorded and transcribed by the authors.
- 57 Jessica Dorsey et al., 'Bridging Understandings of the Military AI Lifecycle: A Transdisciplinary Socio-Technical Approach to Governance', *Opinio Juris*, 12 June 2026, at: <https://opiniojuris.org/2026/06/12/bridging-understandings-of-the-lifecycle-interdisciplinary-approaches-to-military-ai-governance>.
- 58 IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*.
- 59 Jessica Dorsey et al., 'Designing for Responsibility: Outlining an Interdisciplinary Lifecycle Framework Research Agenda for AI-Enabled Weapons and Decision-Support Systems', conference paper under review, 2026; Jonathan Kwik, 'Iterative Assessment for Military Artificial Intelligence (AI) Systems', in Bérénice Boutin et al. (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (T.M.C. Asser Press, 2026).
- 60 The AutoPractices Project, *Strengthening Human Agency in the Military Domain*; Ariel Conn and Ingvild Bode, 'Establishing Human Responsibility and Accountability at Early Stages of the Lifecycle for AI-Based Defence Systems', *Ethics and Information Technology* 27, no. 51 (2025), at: <https://doi.org/10.1007/s10676-025-09862-1>.
- 61 Ali Mostashari et al., 'Developing a Stakeholder-Assisted Agile CONOPS Development Process', *Systems Engineering* 15, no. 1 (2012): 1, at: <https://doi.org/10.1002/sys.20190>.

- 62 Klaudia Klonowska, 'AI-Enabled Decision Support Systems as Means and Methods of Warfare', *Legal Review of Weapons Information Portal*, 1 October 2025, at: <https://legalreviewportal.org/technologies/decision-support-systems>.
- 63 Ibid.; Tobias Vestner and Altea Rossi, 'Legal Reviews of War Algorithms', *International Law Studies* 97 (2021): 509–555.
- 64 See, for example, Department of National Defence and Canadian Armed Forces, *Artificial Intelligence Strategy* (Government of Canada, 2024), at: www.canada.ca/content/dam/dnd-mdn/documents/reports/ai-ia/dndcaf-ai-strategy.pdf.
- 65 Zena Assaad, 'A Proposed Risk Categorisation Model for Human-Machine Teaming', paper presented at EICS '22: Engineering Interactive Computing Systems conference, 21 June 2022, *CEUR Workshop Proceedings* (2022), at: <https://ceur-ws.org/Vol-3404/paper11.pdf>; Alexander Blanchard et al., 'A Risk-Based Regulatory Approach to Autonomous Weapon Systems', *Digital Society* 4, no. 23 (2025), at: <https://doi.org/10.1007/s44206-025-00181-y>; Jack Shanahan, 'Symposium on Military AI and the Law of Armed Conflict: A Risk Framework for AI-Enabled Military Systems', *Opinio Juris*, 1 April 2024, at: <https://opinio-juris.org/2024/04/01/symposium-on-military-ai-and-the-law-of-armed-conflict-a-risk-framework-for-ai-enabled-military-systems>.
- 66 International Committee of the Red Cross, *Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain*, p. 6.
- 67 We are grateful to Neil Renic for drawing our closer attention to these points.
- 68 Renic and Schwarz, 'Crimes of Dispassion'; Neil Renic, 'Tragic Reflection, Political Wisdom, and the Future of Algorithmic War', *Australian Journal of International Affairs* 78, no. 2 (2024): 247–256, at: <https://doi.org/10.1080/10357718.2024.2328299>.
- 69 Yuval Abraham, "'Lavender': The AI Machine Directing Israel's Bombing Spree in Gaza", *+972 Magazine*, 3 April 2024, at: www.972mag.com/lavender-ai-israeli-army-gaza.
- 70 Kozlovski, 'Meaningful Human Control over AI Military Decision Support Systems'.
- 71 International Committee of the Red Cross, *Submission to the United Nations Secretary-General on Artificial Intelligence in the Military Domain*.
- 72 Netta Goussac and Vincent Boulanin, *Responsible Procurement of Military Artificial Intelligence* (Stockholm International Peace Research Institute, 2026), at: <https://doi.org/10.55163/YOLG1827>.
- 73 Kevin LaGrandeur, 'The Consequences of AI Hype', *AI and Ethics* 4, no. 3 (2024): 653–656, at: <https://doi.org/10.1007/s43681-023-00352-y>.
- 74 Neil Renic, 'AI-Optimized Violence and the Suffocation of Moral and Political Wisdom', *Cambridge Forum on AI: Law and Governance* 1 (2025): 8, at: <https://doi.org/10.1017/cfl.2025.10025>.
- 75 Stephanie M Tully et al., 'Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity', *Journal of Marketing* 89, no. 5 (2025): 1–20, at: <https://doi.org/10.1177/00222429251314491>.
- 76 Alexander Blanchard and Laura Bruun, *Bias in Military Artificial Intelligence* (Stockholm International Peace Research Institute, December 2024), at: <https://doi.org/10.55163/CJFT9557>; Laura Bruun and Marta Bo, *Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law* (Stockholm International Peace Research Institute, 2025); Kwik, 'Digital Yes-Men'.
- 77 Holland Michel, *Decisions, Decisions, Decisions*.
- 78 Jeffrey M Bradshaw et al., 'The Seven Deadly Myths of "Autonomous Systems"', *IEEE Intelligent Systems* 28, no. 3 (2013): 54–61; John K Hawley, *Patriot Wars: Automation and the Patriot Air and Missile Defense System* (Center for New American Security, 2017), at: www.cnas.org/publications/reports/patriot-wars; Ingvild Bode and Tom FA Watts, *Meaning-Less Human Control: Lessons from Air Defence Systems on Meaningful Human Control for the Debate on AWS* (Drone Wars UK and Center for War Studies, 2021), at: <https://dronewars.net/2021/02/19/meaning-less-human-control-lessons-from-air-defence-systems-for-lethal-autonomous-weapons>.
- 79 Davidovic, 'On the Purpose of Meaningful Human Control of AI', p. 2.
- 80 Nataliya Kosmyrna et al., 'Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task', arXiv:2506.08872, *arXiv*, 10 June 2025, at: <https://doi.org/10.48550/arXiv.2506.08872>; Michael Gerlich, 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking', *Societies* 15, no. 1 (2025): 6, at: <https://doi.org/10.3390/soc15010006>.

- 81 Chan et al., 'Harms from Increasingly Agentic Algorithmic Systems'.
- 82 On the various roles of industry in military AI development, see Alexander Blanchard et al., 'Mapping the Military AI Industry', *Stockholm International Peace Research Institute* (website), 23 April 2026, at: www.sipri.org/commentary/topical-background/2026/mapping-military-ai-industry.
- 83 The Global Commission on Responsible Artificial Intelligence in the Military Domain, *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (The Hague Centre for Strategic Studies, 2025), p. 57, at: <https://hcass.nl/wp-content/uploads/2025/09/GC-REALM-Strategic-Guidance-Report-Final-WEB.pdf>.
- 84 Klonowska and Woodcock, 'Rhetoric and Regulation'; Yiokasti Mouratidi, 'Armed Groups and International Law—Beyond Compliance Symposium: Before Compliance—Ex-Ante Decision Making in the Design of Military AI Capabilities and the Need to Unravel the Private Technology Sector's Role', *Armed Groups and International Law*, 23 September 2025, at: www.armedgroups-internationallaw.org/2025/09/23/beyond-compliance-consortium-before-compliance-ex-ante-decision-making-in-the-design-of-military-ai-capabilities-and-the-need-to-unravel-the-private-technology-sectors-role.
- 85 Ioana Puscas, *Human-Machine Interfaces in Autonomous Weapon Systems* (United Nations Institute for Disarmament Research, 2022), at: <https://unidir.org/publication/human-machine-interfaces-in-autonomous-weapon-systems>.
- 86 Australia, *Australia's System of Control and Applications for Autonomous Weapon Systems*, para. 13.
- 87 Bode et al., 'Ensuring the Exercise of Human Agency in AI-Based Military Systems'.
- 88 Anna Nadibaidze et al., *AI in Military Decision Support Systems: A Review of Developments and Debates* (Center for War Studies, 2024), at: www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates.
- 89 Australia, *Australia's System of Control and Applications for Autonomous Weapon Systems*.
- 90 HW Meerveld et al., 'The Irresponsibility of Not Using AI in the Military', *Ethics and Information Technology* 25, no. 1 (2023): 14, at: <https://doi.org/10.1007/s10676-023-09683-0>; Emelia S Probasco et al., *AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks* (Center for Security and Emerging Technology, 2025), at: <https://cset.georgetown.edu/publication/ai-for-military-decision-making>.
- 91 Department of Defence, *2026 National Defence Strategy*; Department of Defence, *2026 Integrated Investment Program*, p. 79.
- 92 Ibid., p. 79.
- 93 Wood, 'Autonomous and AI-Enabled Systems', p. 10.
- 94 Ibid.
- 95 Kozlovski, 'Meaningful Human Control over AI Military Decision Support Systems', p. 138.
- 96 Trabucco, 'AI-Enabled Autonomous Weapons and Human Control. Part I', p. 547.
- 97 Australian Delegation at the Informal Exchanges on AI in the Military Domain and Its Implications for International Peace and Security, United Nations in Geneva, 15 June 2026. Quote recorded and transcribed by the authors.

/ RESPONSIBLE PROCUREMENT OF MILITARY ARTIFICIAL INTELLIGENCE IN AUSTRALIA: A WINDOW OF OPPORTUNITY

Netta Goussac and Vincent Boulanin

Introduction

States are rushing to adopt artificial intelligence (AI) in the military domain. Australia's 2026 National Defence Strategy, for example, states that 'Defence will demonstrate, develop, certify and integrate artificial intelligence into capability safely and rapidly'.¹ Converging factors are driving this urgency. Strategic competition between the major powers has intensified, with the United States of America, China and Russia each investing heavily in military AI as a source of competitive advantage.² Recent conflicts have reinforced the perception—first expressed by the US in its 2016 'third offset strategy'³—that AI will be central to the conduct of modern warfare. In the 2020 Nagorno-Karabakh conflict, AI-enabled loitering munitions played a significant role in Azerbaijan's victory; since 2022, Ukraine's response to Russia's invasion of its territory relied on innovative uses of robotics and AI; Israel has relied on AI extensively to plan and conduct military operations in its conflict with Hamas in Gaza since 2023; AI was used by Israel and the US to identify targets and compress decision timelines in the 2026 conflict with Iran.⁴

How states should pursue the adoption of AI capabilities in the military domain remains, however, a challenging question, as they face two difficult and interrelated issues when it comes to the procurement of such capabilities.

The first concerns the need for responsibility. States need to ensure the AI capabilities they procure are safe, secure, trustworthy, lawful and aligned with the political commitments they have made around responsible military AI at the national or international level.

The second pertains to the need for speed. In a deteriorating geopolitical environment many states are under pressure to accelerate their adoption of AI in the military.⁵ AI is perceived as offering an indispensable strategic and operational advantage. This narrative is echoed and amplified by industry actors, who position AI capabilities as essential to extracting actionable insights from ever-growing flows of intelligence, surveillance and reconnaissance data, and deploying more autonomous weapon systems.⁶ A chorus of voices from militaries and industry alike have called for reform of military procurement processes to allow more rapid acquisition and fielding of AI capabilities. States are still grappling with finding a balance between their desire to streamline procurement pathways to facilitate the adoption of AI and the commitment to preserve the elements that ensure such capabilities will be procured—and ultimately used—responsibly.

This paper explores how Australia could reconcile the perceived need for speed with the need for responsibility in the procurement of military AI capabilities.

Australia's planned reform of procurement is in part driven by the desire to accelerate uptake of new technologies. But it also provides a window of opportunity for Australia to implement its high-level policy commitments and international legal obligations around responsible military AI.

The paper is primarily based on desk research and interviews that the authors conducted on the responsible procurement of military AI in the context of a project funded by the Department of National Defence of Canada and the Ministry of Defense of the Kingdom of the Netherlands. The paper is also informed by observations they made as part of the authors' direct participation in AI-related policy conversations in Australia and multilateral settings, not least the expert discussion on lethal autonomous weapons systems under the auspices of the United Nations Convention on Certain Conventional Weapons and the Responsible AI in the Military Domain (REAIM) Summits.

The paper is structured as follows. The first section discusses the importance of procurement for the responsible adoption of AI in the military domain, and why states' urge to adapt procurement processes to accelerate the acquisition of military capability represents an important window of opportunity for implementing commitments and obligations around responsible military AI. The second section zooms in on the Australian context and maps out the country's policy priorities on military AI. The final section offers recommendations on how Australia could reform its procurement process to balance the need for speed with the need for responsibility.

Procurement and the Responsible Development and Use of AI in the Military Domain

Responsible AI in the Military Domain: from Principle to Practice

States widely regard AI as critical for the future of their military deterrence and warfighting capability. The reported use of AI in recent conflicts, and most recently the US's and Israel's war with Iran, has only reinforced that. For states that view military AI as critical, the key question is no longer if they should use AI for military purposes but how: what they should adopt, when, and how quickly.

As a recent report published by the UN Secretary-General noted, states broadly agree that rushing to adoption entails risks and that they should therefore seek to adopt AI responsibly—that is, with due regard for legal obligations, safety and security, as well as ethical considerations. This has led to the emergence over the past few years of multiple policy initiatives at the national and multilateral levels related to responsible AI. France, the US, the UK, Japan, Singapore and Australia have developed national principles on the development and use of military AI. NATO and the EU (via the European Parliament) have also set out principles to guide their member states. The Netherlands and the US have initiated two separate multilateral forums for states to discuss international principles—the REAIM Summit, and the US-led Political Declaration on Responsible Military Use of AI and Autonomy. These processes have led to the formulation of policy principles and commitments that many states around the world subscribe to. For instance, the Blueprint for Action adopted at the REAIM Summit in 2024 was endorsed by 60 states. The principles and policy commitments that came out of these various initiatives vary in formulation and purpose, but these commitments commonly touch on a core set of themes: (1) equitability and bias mitigation, (2) traceability and explainability, (3) reliability and security, (4) accountability and (5) governability.⁷

A key task facing states now is to elaborate on how these very general principles are to be implemented in practice. This question was the core focus of the February 2026 REAIM Summit. A recurring talking point at the summit was that many of the principles require action in the design and development of AI systems. Surprisingly, little was said about the role of procurement, although it is one of the mechanisms through which these actions can be put into motion.

Procurement as a Key Intervention Point

Military procurement's primary purpose is to enable capability adoption by bridging a military's strategic needs and its operational capabilities. The strategic and security imperatives, as well as the complexity of some military capabilities, mean that many countries have introduced specific procurement processes dedicated to the military

domain.⁸ Military procurement is a formal and often sequential process that typically involves identifying a capability gap, formulating a technological need, soliciting and evaluating proposals from suppliers, and awarding and managing contracts, as well as testing and delivering capabilities to the armed forces.

In practice, military procurement processes are more than a process of acquiring capabilities; they also function as mechanisms for states to implement policy commitments and national and international legal obligations, including regarding responsible military AI.

Procurement has barely been discussed in policy debates on the responsible adoption of military AI, although it is an obvious vehicle for states to implement their commitments in this domain.⁹ The procurement process provides many intervention points, from assessment of operational requirement to technology evaluation, to delivery management and review. It also provides a framework for states to think holistically about how their various commitments around responsibility can be tied together. These include their commitments to make military AI by design safe, secure, ethical and lawful; conduct legal reviews; conduct robust testing and evaluation; and ensure that end users are trained to understand how a capability may and may not be used.

Procurement Reform as a Window of Opportunity

The fact that procurement has not received much attention in debates on responsible military AI is all the more surprising as many states, the US in particular, have been vocal over the past decade about their intention to reform their military procurement processes to facilitate the adoption of AI and emerging technologies.¹⁰ The common rationale for such reforms is threefold.

First, the pace of the procurement process is too slow. Procurement processes can take years, if not decades, largely due to an extensive bureaucratic decision-making process. The pace and model of military procurement stand in sharp contrast with the way in which and the speed at which AI products are typically developed and deployed in the civilian sector. AI product development cycles are measured in weeks or months.

Second, traditional procurement typically follows a waterfall model, moving sequentially through formal stages.¹¹ This linear approach assumes that capability requirements remain static throughout. By contrast, the development of AI capabilities—whether in the military or civilian domain—requires constant iteration; otherwise there is a risk that the technology once deployed is already obsolete. The inadequacy of waterfall methodologies for software systems was identified as early as the 1980s, and procurement agencies have been working to adapt their process accordingly, but with mixed success. It is not uncommon that military equipment enters into service while relying on software that is already outdated.

Third, military procurement is highly system/capability-centric: it is structured around the procurement of a specific weapon (e.g. type of missile), platform (e.g. aircraft) or system (e.g. air defence system). Critical AI capabilities may, however, operate as part of a system-of-systems involving various types of sensors, computing and communication infrastructure, and actuators (whether weapons or other military systems). This adds complexity, as it requires procurement agencies to think more holistically and consider many forms of technical and socio-technical interdependencies when articulating requirements for the design of the AI capability and the assurance process.

Fourth, the process and methods that procurement agencies typically use to verify that the technology they procure is reliable, safe and secure are not adequate for the latest advances of AI.¹² These assurance methods were initially designed for hardware and software that exhibit deterministic behaviour—that is, given the same inputs and conditions, they produce predictable, measurable, replicable and verifiable outputs.¹³ AI capabilities, which are now based on machine learning, are, in contrast, non-deterministic. Formal methods of verification do not (yet) work well with them. The reliability, safety, and security of machine learning systems need to be evaluated through other means such as simulations and operational testing. Such empirical tests and evaluations are useful but can only provide a partial picture of how systems perform and how they may fail.

2. The Australian Context: Clear Focus on Speed with Less Guidance on Responsibility

Converging developments render the topic of military AI procurement timely and significant in the Australian context, where accelerating military acquisition of AI is on the agenda but there are relatively few (openly available) guidelines about responsible development and use of military AI. Relevant developments include Australia's participation in AUKUS Pillar II, Australia's engagement in multilateral processes on responsible AI and, more recently, the publication of the Department of Defence's Policy Settings for Responsible Use of Artificial Intelligence in Defence.

The Need for Speed: AUKUS Pillar II and the Advanced Strategic Capabilities Accelerator

Australia's participation in AUKUS Pillar II and its Advanced Strategic Capabilities Accelerator (ASCA) reflects a clear strategic orientation towards accelerated military AI adoption, particularly in the development and fielding of AI-enabled capabilities and autonomous systems.

AUKUS Pillar II addresses six technological areas, including AI and autonomy.¹⁴ A working group is dedicated to each technological area, as well as to two functional areas: innovation (accelerating the countries' respective defence innovation enterprises 'to more rapidly integrate commercial technologies to solve warfighting needs') and information sharing to enable work related to Pillar II.¹⁵ Working groups may engage in a number of activities to accomplish their goals,¹⁶ and much AUKUS-related activity relating to advanced capabilities is classified.¹⁷ Despite this, it is clear that compressing development and acquisition timelines is an ambition of Pillar II. A 2022 fact sheet described early work in the 'AI and autonomy' working group as being focused on 'accelerating adoption, and improving the resilience of, autonomous and AI-enabled systems in contested environments'.¹⁸ The three AUKUS states' defence ministers announced in 2024 that 'the National Armaments Directors from each nation are identifying opportunities to harmonise acquisition processes and reducing barriers to facilitate the accelerated delivery of Pillar II advanced capabilities'.¹⁹

Established in July 2023 following the Defence Strategic Review, ASCA functions as a dedicated mechanism for accelerating the transition of advanced capabilities from development into operational use.²⁰ ASCA's stated mission is to accelerate the development and transition of asymmetric capabilities to the ADF through innovation, in order to meet Defence priority needs.²¹ ASCA does not develop capabilities itself; rather, it connects and streamlines the defence innovation system to drive capability development and acquisition pathways at speed,²² coalescing with Australia's longstanding concern about sovereign defence industrial capability.²³ ASCA is pursuing time-limited 'missions' that respond to priorities with minimum viable capabilities.²⁴ A priority for the program is 'trusted autonomy',²⁵ and these discrete missions encompass AI capabilities, such as 'Ghost Shark—Mission Zero', which is an underwater autonomous vehicle program,²⁶ and 'Mission Syracuse' to support a counter small uncrewed aerial system capability.²⁷ Alongside these discrete missions, ASCA has established a dedicated program for longer-horizon AI development. The Emerging and Disruptive Technologies (EDT) Decision Advantage program is oriented towards 'discovery, new knowledge, concepts and prototypes' in areas including automated data integration, machine reasoning, AI agents and AI orchestration.²⁸

The Need for Responsibility: International Commitments and National Policy

When it comes to the need for responsibility, Australia's orientation has remained more general. Australia has been an active participant in the principal international forums on responsible military AI. It participated in the three REAIM summits—at The Hague in 2023, Seoul in 2024 and A Coruña in 2026—and has endorsed the outcome documents of each summit.²⁹ Australia has also endorsed the US-led Political Declaration on Responsible Military Use of AI and Autonomy (the Political Declaration)—a framework

launched in 2023 to build international consensus around responsible behaviour and guide states' development and use of military AI.³⁰ In its submission to the UN Secretary-General's report on AI in the military domain, Australia affirmed that the military use of AI should be ethical, responsible, and enhance international security, and must comply with applicable international law, including international humanitarian law (IHL) and international human rights law, throughout the capability lifecycle.³¹

Yet at the national level, Australia's measures to give effect to these obligations and commitments are less defined. Australia has published only one policy document on responsible AI in the military domain (others may exist but are not available publicly). The Policy Settings for Responsible Use of Artificial Intelligence in Defence, published in March 2026, are intended to implement Australia's international commitments and its obligations under international law. The settings include three broad requirements relating to the use of AI in the Australian Defence Force: lawfulness, adherence to values-based principles, and proportionate controls.³² The document outlines control measures, such as testing and evaluation, post-incident/impact reviews, and reporting and oversight; and control models, such as direct human oversight and supervision, human-machine teaming and machine-machine teaming, that it says will be applied based on levels of risk and the characteristics of specific AI technologies.³³ The document is intended to fill a carve-out for the Defence portfolio in the Australian Government's 2024 Policy for the Responsible Use of AI in Government but is considerably less detailed than this antecedent.³⁴

The result is that Australia currently occupies an awkward intermediate position. It has made important commitments to responsible military AI and reaffirmed relevant obligations under international law. But it has not articulated the institutional architecture for translating these commitments into practice, including in the procurement phase.

How Australia manages this tension will be consequential for its relationship with partner states like the UK and the US. The UK has more elaborate national policies on responsible military AI, including the 2022 Defence AI Strategy and the 2024 Directive on Dependable AI in Defence.³⁵ The US has similar—if more detailed—policies, developed even earlier, including the Department of Defense Directive 3000.09, ethical principles for AI, and Responsible AI Strategy and Implementation Pathway.³⁶ But in January 2026, the US announced its AI strategy for the (now) Department of War. This strategy shifted emphasis to speed and lethality, mandating 'any lawful use' of AI and directing removal of barriers to rapid deployment. Other developments reinforce the uncertainty about the US position with respect to responsible military AI and the signal it sends to allies and partners. This includes the absence of progress in the process led by the US Department of State to elaborate implementation measures under the Political Declaration. Moreover, a dispute with Anthropic, a supplier, over the permissible use of its AI model by the US military saw the Department of War threaten to terminate the company's contract and

designate it a ‘supply-chain risk’ after Anthropic refused to allow its Claude model to be used for fully autonomous lethal targeting and domestic mass surveillance.³⁷

This tension between strategic orientation towards rapid procurement of military AI and high-level commitments to responsible military AI is also consequential for the significant reforms Australia is undertaking of its defence procurement architecture. This moment offers an opportunity for Australia to design and put in place the institutional architecture and processes necessary to give effect to its commitments around responsible AI, thereby developing a national approach and clarifying its position in relation to its allies. The next section offers a few recommendations to that effect.

3. Seizing the Opportunity to Embed Responsible Military AI in a New Procurement Environment

In 2025, Australia announced significant defence procurement reforms, including the creation of a Defence Delivery Agency (DDA) that consolidates the functions and budgets of Australia’s principal procurement organisations—the Capability Acquisition and Sustainment Group), the Guided Weapons and Explosive Ordnance Enterprise and the Naval Shipbuilding and Sustainment Group (though notably not the Australian Submarine Agency).³⁸ The drivers for this reform are familiar: long delays, cost overruns and blurred lines of responsibility.³⁹ The establishment of the DDA reflects a recognition that existing acquisition processes do not deliver capability at the pace, scale and price that Australia demands. With respect to technologies such as military AI, the 2024 National Defence Strategy stated that Australia is ‘reforming its approach to contracting with defence industry’ and developing ‘[m]ore flexible contracting mechanisms ... to support agile and developmental projects where technology is rapidly changing or where the ADF’s requirements cannot be fully known’.⁴⁰

This moment offers an opportunity to implement legal obligations and operationalise high-level policy commitments to responsible development and use of military AI through procurement processes.

Procurement can serve as a mechanism for implementing responsible military AI, but only if deliberately structured to do so. Procurement represents a critical juncture, prior to the use of military AI capabilities, where legal obligations and high-level policy commitments can be operationalised. A confluence of factors—sharpened focus on accelerated adoption through AUKUS Pillar II and ASCA, new responsible military AI policy settings and structural procurement reform—creates a policy opportunity for Australia. The following recommendations are aimed at ensuring that the institutional architecture for procuring military AI technologies at pace is built with responsible procurement in mind.

Connect Responsible Development and Use of Military AI to Responsible Procurement

A critical first step in giving effect to legal obligations and policy commitments regarding responsible military AI is to ensure that policies, procedures and practices for military procurement explicitly refer to any national laws, policies and commitments related to responsible development and use of military AI.⁴¹ Without explicit cross-referencing, responsible AI principles risk remaining siloed within governance documents that are institutionally distant from procurement decision-makers.

For Australia, each component of the suite of new policy and procedural documents that will necessarily accompany the establishment of the DDA—including its founding instruments, manuals, templates and oversight frameworks—should make explicit reference to Australia’s commitments under IHL, Defence’s policy settings, and outcome documents of REAIM summits and the Group of Governmental Experts on Lethal Autonomous Weapons Systems, as well as the US-led Political Declaration. Similarly, the terms of reference and operational procedures being developed for ASCA missions involving AI-enabled capabilities should be reviewed against these commitments. Australia’s policy settings for responsible military AI set the example, expressly referencing Australia’s commitments to REAIM outcome documents, the Political Declaration and international law.

Independent Capacity to Test Supplier Claims Within the DDA

Australia should ensure that the design of the DDA explicitly includes independent testing, evaluation, validation and verification (TEVV) capacity for AI-enabled capabilities, whether in-house or using trusted third-party providers. Legal obligations and policy commitments regarding responsible military AI require procurement authorities to maintain independent capacity to test supplier claims.⁴² Assessing supplier claims requires the technical literacy to interrogate whether and why a military AI capability is needed, and to evaluate what suppliers assert about the performance and reliability of their systems. And the pace at which Australia seeks to transition from prototyping to fielding AI capabilities makes it even more important to ensure that the DDA is designed and resourced at the outset with the technical expertise and institutional authority to conduct or commission independent TEVV rather than relying on supplier-provided assurance.

Integrate Legal Review Requirements into ASCA and DDA Processes

ASCA missions and DDA processes involving AI-enabled capabilities involve precisely the kinds of activities that trigger the need for a legal review under Article 36 of Additional Protocol I to the Geneva Conventions. The obligation to conduct such legal reviews is addressed prominently in Australia's policy settings for responsible military AI, reflecting its longstanding focus on the importance of such reviews in the governance of autonomous weapon systems and other applications of AI in the military domain.⁴³ Legal review requirements should be explicitly embedded within ASCA's mission governance framework and within the DDA acquisition procedures.

In doing so, Australia should address the challenges of reviewing military AI capabilities. The integration of AI into military systems introduces distinctive characteristics that challenge traditional legal review processes and methodologies.⁴⁴ Critically, the scope of Article 36 is limited to the review of new weapons, means and methods of warfare.⁴⁵ This raises interpretative questions about which military AI capabilities to review and when, which could be clarified in new procedures.⁴⁶ For example, even where AI capabilities are subject to legal reviews, the iterative and fast-paced approach taken to the acquisition of some AI-enabled military capabilities, and the challenges of reliability and assurance associated with opaque and brittle systems, potentially complicate the linear process of legal reviews within the procurement process.⁴⁷ Reform of procurement offers an opportunity to adapt legal review processes to these challenges, including by subjecting all military AI capabilities to legal review, regardless of whether they fall within the Article 36 category of a new weapon, means or method of warfare.

Develop and Publish Guidance for Industry on Responsible Military AI

Australia should develop and publish documents articulating clear expectations for suppliers of military AI capabilities, addressing the technical parameters and performance standards necessary to implement its policy settings, including the values of accountability; bias and harm mitigation; explainability; human-centrism; and reliability and security.⁴⁸

Australia's approach to military AI procurement—including ASCA and the DDA—draws in a wide range of industry actors, including Australian primes, small and medium-sized businesses, and research organisations. Many of these actors will not have established internal frameworks for translating responsible military AI principles into product specifications. Without such guidance, procurement authorities and suppliers alike face a more complex and lengthier process of pre-contract evaluation, back-and-forth requests for documentation, and uncertainty around requirements, which can put at risk the delivery of the capability on time and on budget.⁴⁹ The implementation measures directed at giving effect to Defence's 2026 policy settings provide a moment to initiate the development of industry-facing guidance.

Take Opportunities to Broaden and Strengthen Responsible Procurement of Military AI with Partners and Neighbours

States adapting their procurement processes are navigating fundamental implementation questions independently, without the benefit of shared learning or common vocabularies. Responsible procurement should be an explicit component of international policy discussions on military AI governance.⁵⁰ Opportunities for such conversations exist at the REAIM summits, at the ongoing workshops among states supporting the US-led Political Declaration and the US-led AI Partnership for Defence, and at any discussion on AI in the military domain under the auspices of the UN General Assembly First Committee.⁵¹ Australia is well positioned to contribute to addressing this gap, both within its principal security partnership and across its broader region.

Within AUKUS, the existing commitment to harmonising capability development and acquisition systems creates an opportunity for developing common responsible AI procurement standards among Australia, the US and the UK. Australia should advocate for a responsible AI procurement component within that harmonisation agenda, such as shared TEVV frameworks and common supplier expectations across the three partners.

Australia also maintains defence relationships with a range of regional partners such as Indonesia, Japan, the Philippines, the Republic of Korea, Singapore and Thailand, who are themselves beginning to procure AI-enabled military capabilities.⁵² Australia is well placed to work with regional partners to discuss the development of governance frameworks and strengthening of procurement institutions.⁵³

Conclusion

The rapid adoption of AI in the military domain should not come at the expense of responsibility. Procurement is a juncture at which policy commitments and legal obligations can be given practical effect. Australia has made commitments to responsible military AI, and is accelerating adoption through AUKUS Pillar II and ASCA. Forthcoming procurement reforms in Defence provide the opportunity to connect commitments and obligations to actual practice.

The recommendations set out in this paper are aimed at ensuring that new procurement architecture is built with responsibility in mind from the outset. Embedding responsible AI principles at this stage will avoid the need for retrofitting them—a slower, costlier and less effective process. Beyond the national level, Australia has an opportunity to shape how responsible procurement is understood among AUKUS partners and across the region—contributing not only to its own practice but to the emerging frameworks for international governance of military AI.

About the Authors

Netta Goussac is a Senior Researcher in the Governance of Artificial Intelligence Programme at SIPRI. Netta has worked as an international lawyer for over a decade, including as Special Counsel for Lexbridge, for the International Committee of the Red Cross and the Australian Government's Office of International Law, and as a lecturer at the Australian National University. Netta has particular expertise in legal frameworks related to the development, acquisition and transfer of weapons. Netta has provided legal and policy advice and undertaken research related to new technologies of warfare, including autonomous weapons, military applications of artificial intelligence and cyber and space security. Since 2017, Netta has participated in the UN's Group of Governmental Experts on lethal autonomous weapon systems. From 2010 to 2013, she participated in the negotiation of the 2013 Arms Trade Treaty, and subsequently worked to promote universal adherence to its standards.

Dr Vincent Boulanin is a Senior Researcher and Director of the Governance of Artificial Intelligence Programme at SIPRI. He leads SIPRI's research on how to govern the impact of artificial intelligence on international peace and security. He has written extensively on the development, use and control of autonomy in weapon systems and military applications of artificial intelligence. He has briefed the UN Security Council on the impact of emerging technologies on international peace and security and presented several times before the UN Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. He has led collaborative projects with international organisations such as the United Nations Office for Disarmament Affairs (UNODA) and the International Committee of the Red Cross (ICRC), Institute for Electronic and Electric Engineers (IEEE). Before joining SIPRI in 2014, he completed a doctorate in Political Science at École des Hautes en Sciences Sociales in Paris.

ENDNOTES

- 1 Australian Department of Defence, *2026 National Defence Strategy* (Canberra: Commonwealth of Australia, 2026), p. 80.
- 2 See for example discussion in Australian Department of Defence, *2026 National Defence Strategy*, p.16.
- 3 'Remarks by Deputy Secretary Work on Third Offset Strategy', *U.S. Department of Defense* (website), 28 April 2016, at: www.war.gov/News/Speeches/Speech/Article/753482/remarks-by-deputy-secretary-work-on-third-offset-strategy.
- 4 S Shaikh and W Rumbaugh, 'The Air and Missile War in Nagorno-Karabakh: Lessons for the Future of Strike and Defense', *Center for Strategic and International Studies* (CSIS) (website), 8 December 2020; K Bondar, 'Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare', *CSIS*, 20 March 2025; B McKernan and H Davies, 'The Machine Did It Coldly': Israel Used AI to Identify 37 000 Hamas Targets', *The Guardian*, 4 April 2024; R Booth and D Milmo, 'Iran War Heralds Era of AI-Powered Bombing Quicker than 'Speed of Thought'', *The Guardian*, 3 March 2026. See also Ingvid Bode and Tom FA Watts, *Loitering Munitions and Unpredictability: Autonomy in Weapon Systems and Challenges to Human Control* (Odense: Centre for War Studies, University of Southern Denmark, 2023), pp. 52–66, at: https://findresearcher.sdu.dk/ws/portalfiles/portal/231643063/Loitering_Munitions_Unpredictability_WEB.pdf.
- 5 For example, several NATO allies formalised visions for AI adoption in national strategies for their militaries, including France, the UK, Canada and the USA. See French Ministry of Armaments, *L'Intelligence Artificielle au Service de la Défense* [Artificial Intelligence in Support of Defence], Report of the AI Task Force (Ministère des Armes, 2019); British Ministry of Defence (MOD), *Defence Artificial Intelligence Strategy* (London: MOD, June 2022); Canadian Department of National Defence (DND), *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy* (Ottawa: DND, 2024); US Secretary of War, 'Accelerating America's Military AI Dominance', memorandum, 9 January 2026.
- 6 On state narratives, see for example 'Putin: Leader in Artificial Intelligence Will Rule World', *AP News*, 2 September 2017; U.S. Department of Defense (US DOD), *2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (US DOD, February 2019). See also V Boulanin et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (Stockholm: SIPRI, 2020). On the role of industry, see for example S Ghaffray, 'OpenAI Emphasizes China Competition in Pitch to a New Washington', *Financial Post*, 13 January 2025; M Sigalos, 'OpenAI's Altman Warns the US Is Underestimating China's Next-Gen AI Threat', *CNBC*, 18 August 2025; S Elbaum and J Panter, 'DOD's AI Balancing Act', *Council on Foreign Relations* (website), 2 December 2025; J O'Donnell, 'OpenAI's New Defense Contract Completes Its Military Pivot', *MIT Technology Review*, 4 December 2024.
- 7 Netta Goussac and Vincent Boulanin, *Responsible Procurement of Military Artificial Intelligence* (Stockholm: SIPRI, February 2026), pp. 17–18, at: <https://doi.org/10.55163/YOLG1827>.
- 8 OECD, *Brief 23: Public Procurement—Defence Procurement* (OECD, 2011), p. 3, at: www.oecd.org/content/dam/oecd/en/publications/reports/2011/09/defence-procurement_g17a25f1/5js4vmntfz0p-en.pdf.
- 9 An early examination of the issue appears in S Allik et al., 'A Framework for Human Decision-Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications', in *A Framework for Human Decision-Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (IEEE SA, 2024), pp. 42–44.
- 10 See for example G Gentile et al., *A History of the Third Offset, 2014–2018*, RAND Research Report RR-A454-1 (RAND, 2021); K Hicks, US Deputy Secretary of Defense, 'The Urgency to Innovate', keynote address, NDIA Emerging Technologies for Defense Conference, Washington DC, 28 August 2023.
- 11 J Wallin, 'Safe and Effective: Advancing Department of Defense Test and Evaluation for AI and Autonomous Systems', *Center for New American Security* (website), 13 March 2025; JS Gansler, W Lucyshyn and A Spiers, *Using Spiral Development to Reduce Acquisition Cycle Times* (University of Maryland Center for Public Policy and Private Enterprise, 2008).
- 12 C Urban and A Miné, 'A Review of Formal Methods Applied to Machine Learning', arXiv, 21 April 2021.
- 13 RS Panwar, Q Li and JNT Shanahan (eds), *Military Artificial Intelligence Test and Evaluation Model Practices* (INHR, 2024); Wallin, 'Safe and Effective'.

- 14 Government of the United Kingdom, *Fact Sheet: Implementation of the Australia–United Kingdom–United States Partnership (AUKUS)* (Prime Minister’s Office, 2022), at: www.gov.uk/government/publications/implementation-of-the-australia-united-kingdom-united-states-partnership-aukus.
- 15 Ibid.
- 16 Congressional Research Service, *AUKUS Pillar 2 (Advanced Capabilities): Background and Issues for Congress*, R47599 (Washington DC: Congressional Research Service, May 2024), p. 4, at: www.congress.gov/crs-product/R47599.
- 17 U.S. Department of Defense, ‘AUKUS Defense Ministers Meeting Joint Statement’, press release, 1 December 2023, at: www.defense.gov/News/Releases/Release/Article/3604511/aukus-defense-ministers-meeting-joint-statement.
- 18 Government of the United Kingdom, *Fact Sheet: Implementation of the Australia–United Kingdom–United States Partnership (AUKUS)*.
- 19 U.S. Department of Defense, ‘AUKUS Defence Ministers’ Meeting Communiqué’, press release, 26 September 2024, at: www.defense.gov/News/Releases/Release/Article/3918402/aukus-defence-ministers-meeting-communique.
- 20 Australian Department of Defence, ‘Government Announces Most Significant Reshaping of Defence Innovation in Decades to Boost National Security’, media release, 28 April 2023, at www.minister.defence.gov.au/media-releases/2023-04-28/government-announces-most-significant-reshaping-defence-innovation-decades-boost-national-security.
- 21 Ibid.
- 22 ‘Advanced Strategic Capabilities Accelerator’, *Australian Government Directory*, at: www.directory.gov.au/portfolios/defence/department-defence/vice-chief-defence-force/advanced-strategic-capabilities-accelerator (accessed 23 March 2026).
- 23 Tanya Monro, ‘How Can Australia Build Sovereign Capability in AI, and Why Is It That Important?’, *Australian Institute for Machine Learning, University of Adelaide* (website), 28 July 2022, at www.adelaide.edu.au/aiml/news/list/2022/07/28/how-can-australia-build-sovereign-capability-in-ai-and-why-is-it-that.
- 24 ‘Missions’, *Advanced Strategic Capabilities Accelerator* (website), at: www.asca.gov.au/activities/missions (accessed 23 March 2026).
- 25 Australian Department of Defence, ‘Government Announces Most Significant Reshaping of Defence Innovation in Decades to Boost National Security’.
- 26 ‘Ghost Shark—Mission Zero’, *Advanced Strategic Capabilities Accelerator* (website), at: www.asca.gov.au/activities/missions/ghost-shark-mission-zero (accessed 23 March 2026).
- 27 ‘Mission Syracuse’, *Advanced Strategic Capabilities Accelerator* (website), at: www.asca.gov.au/activities/missions/mission-syracuse (accessed 23 March 2026).
- 28 ‘Emerging and Disruptive Technologies: Decision Advantage’, *Advanced Strategic Capabilities Accelerator* (website), at: www.asca.gov.au/activities/emerging-disruptive-technologies/decision-advantage (accessed 23 March 2026); Australian Department of Defence, ‘Harnessing Tech to Drive Faster Decisions across Defence’, news release, 5 January 2026, at: www.defence.gov.au/news-events/releases/2026-01-05/harnessing-tech-drive-faster-decisions-across-defence.
- 29 *REAIM 2023 Call to Action* adopted 16 February 2023 (Government of the Netherlands, 2023); *REAIM Blueprint for Action* adopted 10 September 2024 (Government of the Republic of Korea, 2024); *REAIM 2026 Pathways to Action* adopted 5 February 2026 (Government of Spain, 2026).
- 30 U.S. Department of State, *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* (Bureau of Arms Control, Deterrence and Stability, 2023), at: www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy.
- 31 Australian Government, ‘Australia’s Submission to the United Nations Secretary-General’s Report on Artificial Intelligence in the Military Domain’, response to UN General Assembly Resolution 79/239, 2025, at: [https://docs-library.unoda.org/General_Assembly_First_Committee_-Eightieth_session_\(2025\)/79-239-Australia-en.pdf](https://docs-library.unoda.org/General_Assembly_First_Committee_-Eightieth_session_(2025)/79-239-Australia-en.pdf).

- 32 Australian Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence* (Canberra: Commonwealth of Australia, March 2026), at: www.defence.gov.au/sites/default/files/2026-03/Policy-Settings-for-Responsible-Use-of-Artificial-Intelligence-in-Defence-%5BOFFICIAL%5D.pdf.
- 33 Ibid, p. 9.
- 34 Digital Transformation Agency *Policy for the Responsible Use of AI in Government*, Version 2.0 (Commonwealth of Australia, 2025).
- 35 MOD, *Defence Artificial Intelligence Strategy*; British Ministry of Defence, *Dependable Artificial Intelligence (AI) in Defence—Part 1: Directive*, JSP 936 V1.1 (MOD, 2024).
- 36 U.S. Department of Defense, *Autonomy in Weapon Systems*, Directive 3000.09 (US DOD 2023); U.S. Department of Defense, 'DOD Adopts Ethical Principles for Artificial Intelligence', press release, 24 February 2020; U.S. Department of Defense, Responsible AI Working Council, *U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway* (Washington DC: US DOD, 2022).
- 37 See Jessica Dorsey et al., 'The Pentagon/Anthropic Clash Over Military AI Guardrails', *Opinio Juris*, 26 February 2025; Congressional Research Service, *Pentagon–Anthropic Dispute over Autonomous Weapon Systems: Potential Issues for Congress*, CRS Insight IN12669 (Washington DC: Congressional Research Service, 2026), at: www.congress.gov/crs-product/IN12669. At the time of writing, the dispute remained unresolved and before the courts.
- 38 Richard Marles and Pat Conroy, 'Reforming Defence Capability Development and Delivery', media release, 1 December 2025, at: www.minister.defence.gov.au/media-releases/2025-12-01/reforming-defence-capability-development-delivery.
- 39 Australian Department of Defence, *2023 Defence Strategic Review* (Canberra: Commonwealth of Australia, 2023), para. 12.5, at: www.defence.gov.au/about/reviews-inquiries/defence-strategic-review.
- 40 Australian Department of Defence, *2024 National Defence Strategy* (Canberra: Commonwealth of Australia, 2024), para. 8.4, at: www.defence.gov.au/about/strategic-planning/2024-national-defence-strategy-2024-integrated-investment-program.
- 41 Goussac and Boulanin, *Responsible Procurement of Military Artificial Intelligence*, p. 25.
- 42 Ibid., p. 19.
- 43 Australian Government, *The Australian Article 36 Review Process*, working paper submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons, August 2018, at: [https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_\(2018\)/2018_GGE%2BLAWS_August_Working%2B-paper_Australia.pdf](https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2BLAWS_August_Working%2B-paper_Australia.pdf).
- 44 V Boulanin and M Verbruggen, *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies* (Stockholm: SIPRI, 2017); N Goussac and R Liivoja, 'Legal Review of Military Artificial Intelligence Capabilities', *Articles of War*, 25 August 2025; T Vestner and I Rossi, 'Legal Reviews of War Algorithms', *International Law Studies* 97, no. 509 (2021): 512.
- 45 Article 36, Protocol Additional to the 1949 Geneva Conventions, and Relating to the Protection of Victims of International Armed Conflicts (API), opened for signature 12 December 1977, entered into force 7 December 1978; International Committee of the Red Cross (ICRC), *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare* (Geneva: ICRC: 2006).
- 46 Goussac and Liivoja, 'Legal Review of Military Artificial Intelligence Capabilities'.
- 47 Ibid.
- 48 Australian Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence*, p. 8.
- 49 Goussac and Boulanin, *Responsible Procurement of Military Artificial Intelligence*.
- 50 Ibid.

- 51 On the AI Partnership for Defence, see US DOD, Chief Digital and Artificial Intelligence Office, 'DOD CDAO Holds Eleventh AI Partnership for Defense', press release, 13 August 2025. On discussions under the auspices of the UN General Assembly, see United Nations General Assembly, First Committee, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*, A/C.1/80/L.46 (UN, 2025), para. 10.
- 52 Gatra Priyandita, 'When Regulating AI, ASEAN Should Remain Committed to Responsible Behaviour', *Fulcrum*, 4 June 2024, at: <https://fulcrum.sg/aseanfocus/when-regulating-ai-asean-should-remain-committed-to-responsible-behaviour>; Kendall Stiles, Asfandyar Mir, Jana Puglierin and Karisa Pevehouse, 'Editorial', *Global Policy* 13, no. 3 (2022): 319–321, at: <https://doi.org/10.1111/1758-5899.13106>.
- 53 Japan, the Republic of Korea and Singapore have developed their own responsible military AI frameworks: see Japanese Ministry of Defense, *Guideline for Responsible AI Application in Research and Development of AI-Equipped Defense Systems*, Version 1 (provisional translation) (Ministry of Defense, June 2025); Suon Choi, *South Korean Military-AI Integration: Opportunities and Risks* (Seoul: Asia-Pacific Leadership Network, 2026), at: https://cms.apln.network/wp-content/uploads/2026/03/Choi_March_2026.pdf; Ng Eng Hen, 'Welcome Address at the 3rd Singapore Defence Technology Summit', speech, 12 October 2021, at: www.mindef.gov.sg/news-and-events/latest-releases/12oct21_speech.

/ THE ROLE OF ARTICLE 36 LEGAL REVIEWS ACROSS THE MILITARY AI ASSURANCE LIFECYCLE

Damian Copeland

Introduction

Australia's 2026 National Defence Strategy (NDS) recognises that new technologies, including artificial intelligence (AI) and autonomous weapon systems (AWS), are transforming military capabilities and contributing to the evolution of warfare.¹ This is evidenced by ongoing conflicts in Ukraine, Gaza, Lebanon and Iran, confirming that AI is no longer a novel technology that is peripheral to military capability. It is now embedded in critical military functions including intelligence analysis, sensor fusion, battlespace management, logistics, cyber defence, targeting support and, in some instances, AWS. For many states, including Australia, AI presents a compelling case for military advantage by enabling decision advantage over an adversary; improving the speed and scale of data processing; and enhancing the efficiency of military operations in environments characterised by strategic competition, evolving technological change and finite resources.² Yet the advantages of AI in the military domain also generate legal, ethical and operational risks that must be recognised and mitigated through military AI assurance mechanisms. In the case of the Australian Defence Force (ADF), this is outlined in the Policy Settings for Responsible Use of Artificial Intelligence in Defence (Responsible AI Policy), released by the Department of Defence (Defence) in March 2026, which centres on the principles of lawfulness, adherence to value-based principles and proportionate controls.³

It is widely recognised that AI systems are data dependent and context sensitive.⁴ Their behaviour may vary according to training data, model architecture, integration choices and operational environments.⁵ As a result, their outputs may be predictable and accurate in some conditions and unreliable in others. They may also shape human judgement in ways that are insufficiently visible to operators or commanders.⁶ From a

legal risk perspective, rather than enhancing international humanitarian law (IHL) compliance, an AI system may risk reducing compliance through opaque outputs that increase risks of civilian harm.⁷

These characteristics are especially significant in high-risk applications such as decision support and targeting. While risks are widely recognised in AWS that directly apply kinetic force, they are less well understood where AI materially influences human decision-making that bears upon use-of-force decisions. Such decisions are governed by IHL rules including distinction, proportionality and precautions in attack, and include target development and identification, intelligence assessments relevant to attack decisions, and defensive responses under severe time pressure. However, misplaced reliance may have grave humanitarian or strategic consequences. AI-enabled decision-support (AI-DSS) systems may therefore raise legal, ethical and safety questions comparable in seriousness to those associated with AWS, notwithstanding the absence of a direct weapon function. Such AI use raises the key question of how militaries can assure high-risk AI applications across their lifecycle and what role national article 36 processes play in military AI assurance.

To answer this question, it is useful to consider whether such capabilities fall within a state's article 36 legal review obligation. As AI is recognised as an enabling technology that does not itself apply force or produce a direct military effect but enables the performance of military functions,⁸ this question is not resolved by merely asking whether an AI capability is a 'weapon'. It also concerns whether the capability, in light of its function, design and foreseeable use, is capable of supporting lawful military conduct.⁹

Article 36 requires a state party to Additional Protocol I:

[i]n the study, development, acquisition or adoption of a new weapon, means or method of warfare, to determine whether its employment would, in some or all circumstances, be prohibited by international law.¹⁰

While the article 36 obligation is arguably satisfied by a single determination of legality, the practical need is broader than a late-stage procurement check. Its language points to review across a sequence of acquisition stages: study, development, acquisition and adoption, and so it follows that article 36 is not deliberately confined to a one-off, late-stage assessment. Rather, it can inform a more continuous, lifecycle review process that tracks the evolution of a capability design and the conditions under which it may be used in armed conflict. The ADF's *Guide to the Legal Review of New Weapons, Means or Methods of Warfare* adopts this approach.¹¹

This broader interpretation of article 36 has become more compelling as states and civil society have begun to articulate broader frameworks for responsible military AI. The Global Commission on Responsible Artificial Intelligence in the Military Domain (GC

REAIM) argues in its 2025 report *Responsible by Design* that responsible military AI assurance must be understood across the lifecycle of a system and should include legal and policy evaluation, risk assessment, testing, human–system integration, monitoring, accountability and review.¹² The report adopts the Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA) AI lifecycle framework as a way of structuring these requirements from the period before system development through to review, reuse and retirement.¹³ Almost in parallel, the Stockholm International Peace Research Institute (SIPRI) 2026 report on responsible procurement of military AI identifies the procurement processes as a critical mechanism by which states can implement high-level political commitments and legal obligations, but only if procurement is deliberately structured to do so.¹⁴ Meanwhile, deliberations in the Convention on Certain Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems (CCW GGE LAWS) continue to emphasise, notwithstanding continuing debate on regulatory form, the centrality of human responsibility, accountability and compliance with IHL.¹⁵

These developments are valuable, but they also expose a practical implementation gap. International discussions increasingly converge around high-level principles such as human responsibility, reliability, accountability, traceability and risk mitigation. They are considerably less developed, however, on the questions that must be answered before those principles can be given practical effect: which institutions are responsible for legal and technical standards; what is required at each stage of the lifecycle; what evidence must be generated and assessed before legal conclusions can be drawn; and what legal consequences follow where those requirements are not met. The risk is that ‘responsible military AI’ remains conceptually compelling but practically opaque. For military legal advisers, and especially those engaged in article 36 processes, the question is therefore a practical one: how should legal review processes contribute to a broader assurance framework for military AI, and how should that contribution be organised across the lifecycle of AI-enabled systems?¹⁶

Consistent with Defence’s Responsible AI Policy,¹⁷ this article argues that article 36 legal reviews should be understood as an integral component of the ADF’s broader AI assurance framework. They are not the only assurance mechanism that encompasses technical, organisational, procurement, testing, training and monitoring measures. However, they are an indispensable element of assurance because they provide the legal architecture through which a state determines whether a capability is inherently unlawful, whether it is lawful only under certain conditions, what safeguards are necessary for lawful use and what subsequent changes trigger further review. That role is especially significant for AI-enabled systems, where lawfulness turns not on a single design feature but on the relationship between system function, system limitations, human control arrangements, operational concept and conditions of use.

This article also addresses the relationship between lawful design and lawful use, which creates a socio-technical challenge for effective legal review. Lawfulness cannot be assessed by examining any single technical feature in isolation. A functional assessment of a military AI system requires understanding the full relationship between technical design choices, system performance in operationally realistic conditions, human cognitive and procedural controls, operator training and the institutional context in which the system is deployed. This is a socio-technical challenge as much as it is a legal one. It requires legal advisers to engage with technical evidence in legally meaningful ways, and it requires technical specialists to understand which system properties carry legal significance. Bridging that translation gap between legal requirements and technical specifications, and between technical performance and legal conclusion, is one of the central institutional challenges for military AI governance, and it is one that article 36 processes for AI-enabled systems must be structured to meet.

This article proceeds in five parts. First, it explains what AI assurance in the military domain requires by mapping the development of responsible AI assurance from the international, Commonwealth and Defence policies. It identifies a set of common assurance principles: lawfulness, lifecycle governance, risk-based proportionality, human responsibility and accountability, reliability and security, transparency and explainability, testing and monitoring, and meaningful human control. It then explains why article 36 legal review should be understood as a central component of that broader military AI assurance framework rather than as a narrow or final-stage weapons compliance process. Next, it draws on the author's functional approach to the legal review of autonomous weapon systems to show why function-based legal analysis is particularly well suited to AI-enabled and autonomous capabilities, including AWS and AI-DSS. Finally, it applies that approach across the military AI lifecycle and proposes practical measures for integrating article 36 review into design, development, procurement, testing, human–system integration, operational use, tactical employment, review and retirement. The article concludes that article 36 review is indispensable to giving practical effect to Defence's Responsible AI Policy because it converts broad commitments to lawfulness, accountability and proportionate control into concrete assurance requirements across the lifecycle of military AI.¹⁸

What Does AI Assurance in the Military Look Like?

This article defines AI assurance in the military as a set of legal, technical, organisational and operational measures through which a state obtains justified confidence that AI-enabled capabilities can be designed, developed, acquired, deployed, used, monitored and retired consistently with law, policy and operational requirements.¹⁹ It is not limited to technical validation or model performance testing or satisfied by broad ethical principles

alone. In the military context, AI assurance must address whether an AI system is lawful, reliable, safe, accountable, appropriately governed, operationally effective and subject to context-appropriate human control and judgement across its lifecycle.²⁰

The development of AI assurance in the military context can be traced through a convergence of general AI assurance frameworks, military-specific responsible AI initiatives, and national public-sector AI assurance policies. Although these frameworks differ in legal status, institutional settings and degree of specificity, they increasingly point towards a common architecture of lifecycle assurance that includes risk-based controls, human accountability, transparency, technical robustness, testing and monitoring, and mechanisms for review when systems or circumstances change.

From an Australian perspective, the starting point is the Organisation for Economic Co-operation and Development (OECD) AI Principles, first adopted in 2019 and updated in 2024.²¹ Australia was among the first countries to adopt these principles.²² The OECD AI Principles are recognised as the first intergovernmental standard on AI to 'promote use of AI that is innovative and trustworthy and that respects human rights and democratic values'.²³ They are structured around value-based principles and five recommendations for policymakers. The value-based principles are inclusive growth, sustainable development and wellbeing; human rights and democratic values, including fairness and privacy; transparency and explainability; robustness, security and safety; and accountability. Importantly, the OECD frames these principles across the AI system lifecycle, requiring human agency and oversight, risk management, traceability, responsible disclosure and mechanisms to override, repair or decommission systems where they risk causing harm.

These general principles have influenced the development of Australian AI governance architecture. Australia's AI Ethics Principles, developed by the Department of Industry, Science and Resources, similarly identify human, societal and environmental wellbeing, human-centred values, fairness, privacy protection and security, reliability and safety, transparency and explainability, contestability and accountability as the normative foundation for responsible AI.²⁴ The Australian Government's National Framework for the Assurance of AI in Government, released in June 2024, builds on those principles and identifies assurance as an essential part of AI governance in government, including the development, procurement and deployment of AI.²⁵ It states that assurance enables governments to understand expected benefits, identify risks and mitigations, ensure lawful use, understand whether AI is operating as expected, and demonstrate through evidence that AI use is safe and responsible.²⁶

At the Commonwealth level, this policy architecture has become increasingly operational. In December 2025 the Australian Government's Digital Transformation Agency (DTA) issued its updated Policy for the Responsible Use of AI in Government, which provides

a framework for the accelerated and sustainable adoption of AI by Commonwealth agencies.²⁷ While the policy excludes Defence, it provides insight into what is required of Commonwealth agencies to develop a strategic approach to AI adoption to operationalise responsible AI use, including designating accountability for AI use cases and undertaking risk-based use-case assessments. It is organised around strategy and oversight, preparedness and operations, and AI use-case impact assessment.²⁸ This policy is complemented by the DTA's AI technical standard, which translates responsible AI into practical and technical guidance. It addresses the end-to-end design, development, deployment and use of AI systems and aligns its requirements to an AI lifecycle model.²⁹

Finally, the National AI Centre's guidance for AI adoption provides a useful summary of what operational AI assurance requires in practice. It identifies six essential practices for responsible AI governance: decide who is accountable; understand impacts and plan accordingly; measure and manage risks; share essential information; test and monitor; and maintain human control.³⁰ The implementation guidance is directed to organisations that build or customise AI systems, use AI in complex ways, manage higher-risk use cases, or require stronger controls and oversight. It also emphasises the need for governance frameworks, clear roles, AI registers, supply-chain accountability, system-specific review, clear records, testing, incident monitoring and ongoing improvement.

For Defence these Commonwealth frameworks, despite not applying directly to military AI, are important not because they address the distinctive legal and operational questions raised by military AI but because they identify the basic institutional components of an AI assurance framework. They show that responsible AI governance is not simply a matter of principle. It requires accountable decision-makers, documented use cases, risk assessment and classification, lifecycle records, supplier information, testing evidence, monitoring processes, human oversight arrangements and review mechanisms. Those requirements are even more important in the military domain because AI systems may affect decisions concerning the use of force in targeting, detention, intelligence assessments, force protection and operational planning.

International guidance on military-specific AI assurance has developed most clearly through the REAIM process. In September 2025, the GC REAIM published its *Responsible by Design* report, which provides a comprehensive account of responsible military AI governance.³¹ The report argues that responsibility, including ethics and law, must be integrated 'from the earliest stage of development, through the entire AI system lifecycle' and within the socio-technical institutions in which AI applications are embedded.³² It emphasises that responsible development and use of AI in the military domain requires informed human decisions and context-specific approaches to designing, testing and deploying AI in ways that uphold peace and human dignity.³³ It also stresses that principles must be translated into actionable guidelines and standards for states, militaries and industry.³⁴

In parallel, the REAIM summits have sought to move the discussions on AI assurance towards the concept of ‘responsible AI’ as a practical governance measure of AI-enabled military systems across their lifecycle. It recognises that risks cannot be assessed only at the point of deployment. They arise in capability design, data selection, model development, testing, system integration, human–machine interaction, procurement, operational use, monitoring, updating, reuse and retirement. For military AI, assurance must therefore be continuous, evidence based and institutionally embedded.

The 2026 REAIM Summit, held in A Coruña, Spain, emphasised the need for states to take an operationally focused direction. Its outcome document, the REAIM Pathways to Action, which was signed by Australia, identified the need to translate principles for responsible military AI into concrete measures.³⁵ To achieve this objective, it proposed practical steps for responsible military AI, including legal compliance, sustained human involvement, traceable accountability, risk assessment, documentation of decision-making processes, and cross-regional capacity building.³⁶ Although the document remains non-binding and attracted less support than previous REAIM outcome documents,³⁷ it confirms a broader trend: the responsible military AI agenda is moving from principles to implementation, from declarations to assurance practices, and from general language to concrete lifecycle governance.

Defence’s 2026 Responsible AI Policy follows this international trajectory. This policy states that AI technologies offer opportunities to improve accuracy, efficiency, speed and safety, but that Defence will use AI responsibly and in compliance with Australia’s domestic and international legal obligations.³⁸ Defence further commits to an informed, risk-based approach that preserves individual accountability for AI-enabled decisions and outcomes.³⁹ The policy identifies three obligations: lawfulness; adherence to values-based principles, including individual accountability, human impact, explainability, reliability and security, and mitigation of unintended bias and unintentional harm; and proportionate controls based on risk.⁴⁰

Taken together, these frameworks suggest that military AI assurance should contain at least the following characteristics.

1. **Lawfulness.** Military AI must comply with domestic law and international law. For the ADF, this includes the law of armed conflict, including targeting law, international weapons law, command responsibility, Commonwealth privacy, security and administrative governance obligations.
2. **Lifecycle governance.** AI assurance must begin from development and continue through research, acquisition, testing, deployment, use, monitoring, modification, reuse and retirement. This is consistent with the OECD lifecycle approach, the GC REAIM’s ‘responsible by design’ model, the DTA technical standard and the National AI Centre’s emphasis on testing, monitoring and review.

3. Risk-based assessment. Not all AI systems require the same level of assurance. Higher-risk systems, including those that influence the use of force, targeting, detention, intelligence assessments or command decisions, require higher standards and stringent evidence, human controls, oversight and legal review.
4. Human responsibility and accountability. AI systems may support or shape decisions, but responsibility for military decisions must remain attributable to identifiable human actors. Accountability requires clear roles, records, traceability, command responsibility, supplier responsibility and reviewable decision-making processes.
5. Technical reliability, robustness, safety and security are necessary but not sufficient. Military AI must perform reliably in the conditions in which it is intended to be used, including degraded, contested and adversarial environments. However, technical performance only becomes legally meaningful when assessed against the operational function and legal context in which the system will be employed.
6. Transparency, explainability and information access are central to assurance. Defence must have sufficient information about system design, data, model limitations, confidence measures, update processes and failure modes to assess whether an AI capability can be responsibly and lawfully used. This also has procurement implications: assurance cannot be achieved where suppliers cannot or will not provide the information needed for legal and technical scrutiny.
7. Testing, evaluation, verification, validation and monitoring must be continuous. Pre-deployment testing is necessary but not enough. AI systems may change over time, be used in new environments, or interact with operators in unexpected ways. Assurance must include monitoring, incident reporting, after-action review, revalidation and re-review triggers.
8. Human control and human–system integration must be assessed functionally. The question is not simply whether a human is ‘in the loop’ but whether human involvement is meaningful in light of the system’s function, speed, complexity, interface, confidence signals, operator training and operational context. Human control must be designed, tested, trained and maintained.

These characteristics provide a practical perspective on what AI assurance in the military should look like. It is a lifecycle system of governance in which legal, technical, ethical, operational and organisational controls work together to create justified confidence in responsible AI use. It requires Defence to ask, at each stage of the AI lifecycle: who is accountable; what function does the AI perform; what legal and operational risks arise; what evidence is required; what controls are proportionate; what human judgement is necessary; what testing is adequate; what records must be kept; and what changes require reconsideration?

This framing also clarifies the role of article 36 legal review. Article 36 is not the whole of military AI assurance. It does not replace technical testing, procurement governance, cyber security, human factors analysis, operational doctrine, operator training or command oversight. However, it is the legal mechanism through which the ADF can determine whether AI-enabled and autonomous functions can be used consistently with international law which informs and influences the other components of the AI assurance architecture. It therefore provides the legal scaffolding of an AI assurance framework for ADF development of AWS and AI-DSS by identifying legally significant functions, specifying evidentiary requirements, translating legal risk into design and use constraints, and determining when changes in function, context or performance require further review. This legal review process is the subject of the next section.

Article 36 Legal Review as a Component of Military AI Assurance Architecture

The traditional purpose of article 36 review is well known. It requires a state party to Additional Protocol I⁴¹ to determine whether a new weapon, means or method of warfare would be prohibited by international law in some or all circumstances. The International Committee of the Red Cross (ICRC) *Guide to the Legal Review of New Weapons, Means and Methods of Warfare* emphasises that the review is both substantive and procedural.⁴² Substantively, it asks whether the capability is prohibited by treaty or customary law, whether it is of a nature to cause superfluous injury or unnecessary suffering, whether it is indiscriminate by nature, and whether it would otherwise contravene international law.⁴³ Procedurally, it requires the state to establish national mechanisms capable of obtaining the technical, operational and legal information necessary to make that determination.⁴⁴

In conventional weapon contexts, the ADF often treats article 36 legal review as a point-in-time process conducted once, generally late in the acquisition process.⁴⁵ However, while this approach is appropriate for a static weapon or munition, e.g. a bullet, it is increasingly strained in the AI context. There are several reasons for this. First, the object of review may not be a discrete weapon in the traditional sense. AI as an enabling technology may be embedded in a sensor architecture, software, targeting system or mission-planning tool that does not itself cause an effect but materially shapes the way force is applied. Second, the capability under review may change after initial approval through software updates, changes to data inputs, model replacement, interface redesign or integration with other systems. Third, the legal analysis may depend heavily on socio-technical information that emerges only incrementally through development, testing and operational use.⁴⁶ Finally, most AI governance policies place legality as a central requirement.⁴⁷

For these reasons, it is useful to frame article 36 review as a component of a wider military AI assurance architecture. The Defence Responsible AI Policy recognises this by requiring adherence to three policy requirements: lawfulness, adherence to value-based principles, and proportionate controls.⁴⁸ This understanding is consistent with the *Responsible by Design* report, which treats legal and policy evaluation, testing, monitoring, risk assessment and accountability as interlocking and continuing activities.⁴⁹ An article 36 review fits within that architecture in two ways. First, the legal analysis relies on assurance artefacts, including technical documentation, model performance information, testing outcomes, human–machine interface procedures, incident reporting and operational restrictions. Second, it generates assurance outputs, including legal risk identification, recommendations for use case constraints, compliance red lines, re-review triggers, and findings concerning the conditions under which a capability may or may not be used lawfully.

Seen in this way, article 36 reviews are not just a legal-technical validation. Technical testing cannot answer legal questions on its own. A system may perform well in benchmark conditions and still be incapable of lawful use in a particular operational context. Conversely, legal analysis that is insufficiently informed by technical and operational evidence risks being superficial and inaccurate. The challenge is therefore to ensure that legal review is embedded in a process that allows relevant evidence to be generated, scrutinised and translated into legal judgements at appropriate stages.⁵⁰

That translation challenge is more difficult than it might appear. Legal concepts such as distinction, proportionality and precautions in attack were developed to regulate human conduct in armed conflict. Applying them to AI-enabled capabilities requires translation in both directions: from the technical output to a legal significance; and from the legal requirement to a technical specification. For example, a system's error rate is not in itself a legal concept, but it may become legally significant where errors impact protected persons or objects, are sensitive to adversarial manipulation or occur disproportionately under operational conditions that diverge from the training environment. Conversely, a legal requirement for 'adequate precautions' does not translate automatically into a technical performance threshold. It requires contextual judgement about what level and type of assurance is sufficient given the system's function, the foreseeability of harm and the operational consequences of failure.

Effective legal review of military AI must therefore engage both technical and socio-technical dimensions. This cannot be achieved by legal advisers alone. It requires structured engagement between legal, technical and operational specialists, supported by processes that generate information in legally usable forms. It also requires institutional design: review structures, evidentiary standards, documentation requirements and escalation pathways that enable legal judgements to be made on an adequate evidentiary basis and updated as information accumulates across the lifecycle.

Legal advisers who lack access to relevant technical evidence cannot make meaningful legal determinations. Technical specialists who are not informed about which system properties are legally significant cannot generate the evidence that legal review requires. Addressing that structural gap is necessary for making article 36 reviews effective in the military AI context.

This approach is consistent with the text of article 36 itself. The obligation applies in the ‘study, development, acquisition and adoption’ of a weapon capability.⁵¹ It is therefore difficult to argue that article 36 is detached from early concept design, procurement structuring or later modification. This does not mean that the law demands a full formal review at every step. It means, rather, that states can organise their processes so that legal questions are asked when they can still shape design and use, that re-review is triggered by material changes, and that a final or formal legal determination rests on a body of evidence accumulated over time. This is becoming increasingly dependent on the research and development of military AI by the private sector.

Early article 36 engagement with new technology development is particularly important in the military AI context because legal risk often arises at the intersection of technical design and the contextual use. For example, the legality of an AI-enabled target recognition system may depend not merely on model accuracy in the abstract but on whether it is used as a tool to aid human decisions or for autonomous weapon targeting. Risk may arise due to an operator’s insufficient understanding of a system’s limitations, inaccurate measures of confidence in targeting information, lack of information validation procedures, whether the system is vulnerable to adversarial deception and whether its errors create a foreseeable risk of unlawful targeting. Each of these issues is legal in significance and depends on technical and operational evidence.⁵²

Accordingly, an article 36 process for military AI should include multiple legal assurance steps. As a preliminary issue, a legal review should identify the AI-enabled capability or function that requires legal review. It should identify the evidence required in order to determine compliance with IHL-specific prohibitions and restrictions and general prohibitions. It should translate legal findings into conditions of design, procurement, testing, training and use. Finally, a legal review should determine when the legality of an AI-enabled capability must be reconsidered because a change in design, use context or software may alter the original legal assessment. These functions make article 36 review not a separate legal silo but the legal foundation of a broader AI assurance framework.⁵³

A Functional Approach to the Legal Review of Military AI

The author's work on a functional approach to the legal review of AWS provides a systematic, evidence-based risk-assessment methodology for giving practical effect to article 36 review in the context of high-risk AI use cases. It is designed to enable a legal and contextual analysis of an AI application by focusing on those functions and effects that are regulated by IHL rules such as distinction, proportionality or precautions in attack.⁵⁴

The advantage of a functional approach is that many military AI capabilities do not fit neatly into narrow article 36 concepts of 'weapon' and 'means of warfare'. An AI-DSS may not itself engage a target, but it may influence how a human classifies a person or object as a lawful target or protected. An AI-DSS may also recommend target categories, priorities or weapons employed and so the legal significance of these systems lies in the ability of the human decision-maker to rely on the IHL-regulated outputs, not whether they directly cause a weapon effect. This would allow a legal review to consider a broader range of AI-DSS, for example, a system designed to recommend the status of a captured person during an article 5 tribunal during an international armed conflict.⁵⁵

A functional approach adds substance to the traditional legal review steps and asks a series of practical questions to enable a contextual analysis of the AI system's performance in specific operational circumstances. It aligns with the six essential practices for AI governance recommended by the Australian Government.⁵⁶ For example, what function does the AI perform? Does it identify objects, classify persons, select targets, recommend force options, or directly apply force? What are the system's technical limitations, foreseeable failure modes and confidence boundaries in different operational contexts? What legal risks arise and what safeguards, including human control, are required if the system is to be capable of lawful use?⁵⁷

This functional approach to legal review methodology is consistent with both article 36 and contemporary policy discussions. The ICRC guidance recognises that legal review applies to means and methods of warfare as well as to weapons.⁵⁸ Similarly, the CCW GGE LAWS rolling text identifies the characterisation of systems by reference to their capacity to select and engage targets without intervention by a human user in the execution of those tasks.⁵⁹ The *Responsible by Design* report, although broader in its policy ambition, similarly emphasises the need to understand AI systems across the lifecycle, including their intended role, the allocation of human responsibility and the conditions for trustworthy use.⁶⁰

The most important contribution that the functional approach makes is that it enables a legal determination of AI-enabled systems by focusing only those functions that engage IHL obligations, and mitigates identified risks by recommending levels of human control. So where a system directly selects and engages targets, the legal review will naturally

focus on the system's capacity to comply with article 57 of Additional Protocol I requiring precautions in attack.⁶¹ Where the system instead performs decision-support functions such as object recognition or target recommendation the review will focus on article 48 of Additional Protocol I.⁶² In both cases, the question is whether the capability, in light of its function and operational conditions of use, is capable of being employed consistently with these specific IHL rules.⁶³

The functional approach also helps explain why article 36 legal review must extend across the AI lifecycle. An AI-DSS's function may remain nominally the same while its legal significance changes. A target recognition model may be retrained on new data or it may be integrated into a shorter decision loop with diminished human scrutiny, or a weapon system may shift from a constrained defensive context to a broader offensive role. These changes may alter the reliability, predictability or legal significance of the AI-enabled functions. A functional review is therefore necessarily sensitive to such lifecycle developments.⁶⁴

Finally, the functional approach avoids the mistake of considering human control in binary terms. It recognises that human control is not a static concept and that its legal relevance depends on what risks the use of an AI system raises in any given circumstances. A functional analysis is well suited to assessing these issues by using a risk-based methodology to identify not whether a human is in the loop but whether the allocation of function between human and machine supports lawful conduct in practice.⁶⁵

Article 36 Review Across the AI Lifecycle

The functional approach to the legal review of military AI applies a lifecycle approach to AI assurance. It recognises the need for governance inputs and measures during different stages of the AI system's lifecycle. This directly translates into the nine stages of the IEEE-SA lifecycle framework.⁶⁶

Before System Development

The concept stage of an AI system's development is where many of the most important legal questions should first be identified. The GC REAIM's *Responsible by Design* report places emphasis on identifying intended purpose, constraints, trade-offs, stakeholders and requirements before the technical architecture is locked in.⁶⁷ From an article 36 perspective, this stage should include an initial legal scoping review. The purpose of this exercise is not to deliver a final legal determination but to identify whether the proposed capability is likely to constitute a weapon, means or method of warfare, whether it will perform an AI function that is regulated by IHL rules, and what legal issues will need to be identified as development proceeds.⁶⁸

At this stage, a functional approach is particularly valuable in that it is designed to inform the development of the system by private companies. Private companies developing new military technology may implement internal IHL training or engage external providers to advise on responsible development.⁶⁹ Within Defence research agencies, a military legal adviser may work with system designers, engineers and capability managers to understand the intended operational role of the proposed system and what IHL-regulated functions it will be designed to perform. If the capability is intended to classify objects for target development, the legal adviser should ask how classification will occur, what the data is used and how it is trained to achieve this, what level of confidence is expected, what information validation process is envisaged and how errors may affect downstream decisions. If the system is intended to support autonomous defensive responses, the review should ask what the threat profile is, what target set is anticipated, how tightly the use context can be bounded and what human supervision arrangements are feasible.⁷⁰

The practical legal output of this stage should be an initial or interim legal review that identifies applicable legal rules, key legal risks, evidentiary requirements for further reviews, and timelines or triggers for additional legal consideration. It should also identify whether a particular design or use case is necessary to support lawful use. For example, the interim review may indicate key AI assurance requirements such as that the system must be capable of presenting confidence information; recording decision logs; restricting use to certain environments; or allowing human override at specified points during the planning or conduct of attacks or in cases of doubt. This early legal input is a core assurance function because it shapes the system's requirements before later stages make redesign more difficult and costly.⁷¹

Research and Development

During research and development, article 36 review becomes less theoretical and progressively evidence based as technical information or testing data becomes available. The legal questions identified at the concept stage must now be examined against technical and operational realities. The GC REAIM's *Responsible by Design* report places assurance checks and legal reviews within this stage, alongside data governance, model development and early validation.⁷² This is practical, as a legal review at this point should influence system architecture, training methods, interface design, use cases or models and operational assumptions.

The principal task of legal advisers during development is to ensure that relevant technical and operational evidence is being generated that identifies legal risk. The required evidence will vary according to function but typically includes data provenance; known system limitations; model behaviour under expected conditions and in edge cases; foreseeable failure modes; response to adversarial conditions;

system explainability or interpretability features where relevant; the proposed distribution of cognitive tasks between human users and the AI system; and finally issues of human control and accountability.⁷³

For AWS, this may include evidence concerning how the system detects, tracks, selects and engages targets; how it distinguishes targets from protected persons or objects; what environmental assumptions are necessary for acceptable performance; what fail-safe or abort conditions exist; and whether there are possible gaps in human accountability. For AI-DSS, it may include evidence concerning what legally salient outputs are generated, what uncertainty or confidence measures accompany them, how operators are expected to use those outputs, whether they can interrogate them and whether the system creates foreseeable risks of automation bias or over-reliance.⁷⁴

The functional approach assists by keeping legal analysis tied to AI-enabled or autonomous functionality that engage IHL rules. The question is not whether the model is lawful in the abstract. It is whether its performance is sufficient for the lawful performance of the relevant function in the contexts of its expected use. An AI tool that is acceptable for low-risk maintenance tasks may not be acceptable for high-risk target classification. A system with good average performance but dangerous errors in known edge cases may raise serious legal concerns if used in force-enabling functions. An article 36 review input during system development therefore involves translation between technical capability and legal significance.⁷⁵

A further point of importance is that during research and development, testing may expose the capability as legally unsuitable for one or more of its intended roles in specific circumstances. That possibility should not be treated as a failure of the review process. It is evidence that the review is functioning properly. If development shows that a proposed system cannot achieve performance levels necessary for lawful use or does not allow the exercise of human judgement, the state should modify, constrain or abandon the relevant function. An article 36 review is not merely a mechanism for approving capabilities. It is also a mechanism for identifying when legal constraints require redesign, reduced ambition or prohibition of a proposed use.⁷⁶

Procurement and Acquisition

Procurement and acquisition are legally significant. The Lawful by Design initiative, launched in June 2025, encourages states to include legal review and IHL requirements in their tender requirements and acquisition contracts.⁷⁷ This is reflected in SIPRI's *Responsible Procurement* report, which argues that procurement is a key institutional mechanism through which states can implement political commitments and legal obligations relating to military AI, but only if procurement processes are adapted to require the relevant information, expertise and contractual controls.⁷⁸

From an article 36 perspective, this means procurement must be informed by legal review requirements.⁷⁹ Contracting arrangements should require suppliers to provide data that is sufficient to enable a meaningful legal and technical assessment. That may include system descriptions, model and data documentation, testing results, known limitations, update policies, configuration controls and incident reporting procedures. Procurement should also allow for a state's ability to conduct independent testing to validate supplier claims. A state cannot effectively discharge its article 36 obligations if it acquires an AI capability which is effectively opaque and cannot be independently evaluated.⁸⁰ This creates clear expectations for what militaries require from industry partners during co-development and procurement. At a minimum, procurement documentation should ensure industry provides system documentation sufficient to enable legal assessment.

Where a capability is co-developed with industry rather than commercially procured off the shelf, these requirements should be embedded in the project plan from the outset. Legal advisers must be engaged before technical performance specifications are finalised, not consulted after system architecture has already been established. Legal requirements that are not incorporated into the design from the beginning may be impossible to achieve retroactively. This connects directly to the Lawful by Design framework: legal requirements must shape technical specifications at the earliest stages of co-development, and procurement and co-development contracts are among the most important instruments through which states can operationalise their article 36 obligations in practice.

Procurement also provides an opportunity to translate legal findings into assurance requirements. If legal review has identified the need for decision audit logs, restricted use conditions, retaining human control, human override features or update notification triggers, those requirements should be embedded in procurement documentation, operational procedures and training plans. This is one of the clearest ways in which article 36 review contributes to broader AI assurance. It ensures that legal conclusions are not left as abstract advice but are a design criterion carried into acquisition structures that shape what the state actually receives and how it can lawfully be used.⁸¹

Testing, Evaluation, Verification and Validation

Testing, evaluation, verification and validation (TEVV) processes are critical to article 36 reviews because they provide the empirical data for determining the legality of an AI capability. The GC REAIM report identifies TEVV as a distinct lifecycle stage while also emphasising ongoing testing and monitoring during in-service life.⁸² The significance is that the legality of an AI capability cannot be determined if its testing regime does not examine the specific data, models, use cases and conditions that will impact its lawful use.⁸³

For this reason, the legal adviser should be concerned not only with the existence of testing but with its adequacy and relevance. In the AI context, a narrow focus on performance metrics or laboratory conditions may be misleading. Testing should reflect operationally realistic conditions, including degraded environments, enemy actions, unexpected objects, compressed timelines and human–system interaction effects.⁸⁴ For attack-related functions, legal advisers should pay particular attention to whether testing data enables legal determinations about distinction, precautions and the foreseeability of system error. For decision-support systems, the focus may be on whether operators correctly interpret outputs, whether uncertainty is communicated effectively, and whether error patterns create unacceptable risks in the decisions the system influences.⁸⁵

Testing should also be linked to the conditions needed for legal approval. If lawful use depends on the existence of a specific operational environment, the testing regime must validate performance in that environment and identify the environmental circumstances in which performance becomes unreliable. If approval depends on human verification of certain outputs, testing must examine whether the human users are able to perform the verification role effectively. If use is limited to defensive operations against specified object classes, testing should support those specific constraints. In summary, TEVV should be designed not purely as a technical exercise but also as the evidentiary basis for legal findings and conclusions.⁸⁶

Human–System Integration, Education and Training

The lawful use of military AI will turn on how human beings interact with the system. The GC REAIM report regards human–system integration, education and training as a distinct lifecycle stage.⁸⁷ This reflects the reality that even a technically capable system may be incapable of lawful use if operators do not understand its intended purpose, its performance limitations and when human intervention is required.

The functional approach to legal review applies a risk-based assessment of the proposed human operator's role.⁸⁸ A legal review should consider the design of interfaces, the presentation of confidence or uncertainty information, the timing and pace of decisions, the scope for human intervention and the content of operator training programs. It should also examine known cognitive risks, including automation bias. A human decision-maker who is functionally 'in the loop' but presented with opaque outputs at a pace or in a format that prevents the exercise of human judgement may not provide the kind of human control on which the system's lawfulness is assumed to depend.⁸⁹

Operator training is therefore a critical issue. It is a precondition for lawful use where human judgement is central to legal compliance. An article 36 legal review may identify that a system is capable of lawful use only if operators are trained to understand defined

limitations, verification steps, escalation triggers and circumstances requiring non-use. In this way, article 36 review contributes to assurance by specifying training and human–system integration requirements as part of the legal review report outcomes.⁹⁰

The importance of this stage reflects a broader socio-technical approach to military AI assessment that is increasingly recognised in the academic literature. The AutoPractices Project, led by Professor Ingvild Bode, recognises the need to train operators across the lifecycle to use the interface, continuously cross-check outputs, and identify conditions that might affect or degrade operators' ability to use the interface in compliance with system objectives.⁹¹ Similarly, Assaad and Williams argue that AI decision-support systems introduce multiple layers of complexity into military decision-making, including increased interactivity and nonlinearity, software complexity, and dynamic complexity, requiring appropriate training and education for those operating alongside these systems.⁹² Klonowska and Woodcock have cautioned against legal and regulatory approaches that draw comparisons between human and AI performance in ways that obscure the complexities of human–machine interaction.⁹³ They propose a human–machine interaction framework that reflects the realities of warfare, which aligns directly with the functional approach adopted in this article.

These contributions share a common insight that is directly relevant to article 36 review. The legal assessment of military AI systems can neither be confined to their technical properties nor resolved by merely identifying that a human decision-maker is formally 'in the loop'. Lawful use depends on the quality of the socio-technical interaction between the human operator and the AI system. For article 36 purposes, this means the human–system integration stage must be treated as generating legally relevant evidence to inform a legal review. An article 36 review may need to conclude that a system is capable of lawful use only in conjunction with specific human control arrangements, specific operator training requirements and clear institutional oversight mechanisms that together constitute the conditions of lawful use. Where those conditions cannot be established or maintained, the system is not capable of lawful use, regardless of its technical performance.

Political and Strategic Oversight

Some military AI capabilities may raise strategic and political questions that should not be addressed solely at the technical or tactical level. The GC REAIM report includes political and strategic considerations as a lifecycle stage, reflecting the reality that certain capabilities may affect escalation dynamics, public legitimacy, alliance policy, or the framing of acceptable military risk.⁹⁴ This may be reflected in national military AI policies.⁹⁵ While article 36 review is not a substitute for national policy, it does contribute to this stage by clarifying the boundaries within which national policy choices must operate and by identifying where a capability's lawful use depends on high-level constraints.

To illustrate this point, an AI-enabled defensive system may be legally supportable only within a narrowly bounded defensive envelope and under specific authorization conditions. Alternatively, an AI-DSS may be considered lawful as an analytic tool informing operational planning but not as a targeting tool. A legal review at this stage should therefore inform capability managers of the capability's accepted use cases, its limitations, and the governance measures required to ensure compliance with both legal and national policy requirements.⁹⁶

Operational Command and Control

Once AI systems are deployed, their lawful use is the responsibility of military commanders and system operators. At this stage, referred to as the governance stage,⁹⁷ assurance is not about prospective evaluation but about ensuring lawful use over time. The GC REAIM report emphasises ongoing monitoring, maintenance, updates, interoperability and incident reporting across the AI system lifecycle.⁹⁸ From an article 36 perspective, this means that the system performance, including operational commands, should have mechanisms for tracking, recording incidents, controlling model or software changes, and escalating potential legal concerns.

Consistent with a lifecycle approach, the legal review obligation does not end when the system is fielded. It should continue through mechanisms that monitor whether the assumptions underpinning the original legal determination remain valid. If the operational environment changes materially, if performance degrades, if unexpected failure modes emerge or if operators develop patterns of use that were not anticipated, the legal basis for continued use may need to be reconsidered. The legal advisers deployed in compliance with article 82 of Additional Protocol I will play a role in this ongoing assurance.⁹⁹ This is particularly so for AI-enabled systems that rely on software updates, changing data inputs or integration with other systems.

Tactical Employment

At the tactical level, article 36 review outputs contribute by shaping operational procedures, limitations and conditions of use.¹⁰⁰ System operators should trust that the AWS or AI-DSS issued to them is lawful and ensure that they operate the system lawfully within its approved use cases. That may include restrictions on mission type, environment, target set, confidence thresholds, verification requirements, abort conditions or data source quality.¹⁰¹

The functional approach is again useful because tactical legality depends on the function being performed in a specific operational context. If the system provides a recommendation relevant to target classification, the operator must know what the system limitations are and when independent validation of AI outputs is required. If the system

autonomously tracks and engages incoming threats within a limited envelope, tactical doctrine must reflect the conditions on which legality depends. An article 36 review thus supports tactical assurance by ensuring that legal conclusions are translated into actionable operating parameters.

Review, Reuse and Retirement

The final IEEE lifecycle stage concerns review after use, as well as decisions on reuse, retraining, repurposing or retirement. The GC REAIM report treats this stage as one of further testing, after-action analysis and decisions about sustainment or withdrawal.¹⁰² In the article 36 context, this stage is crucial because military AI may change after initial approval or be employed in new operational contexts that raise new legal risks not considered in previous legal reviews.

As part of a lifecycle assurance process, a legal review should identify clear re-review triggers. These may include substantial software modifications, AI model replacement, retraining on new data, extension to new missions, changes to operating environment assumptions, or evidence of significant unanticipated failure.¹⁰³ After-action reviews and incident reports should be examined for their legal significance. A capability that was lawfully reviewable for one role may not remain so if its function, context or performance profile changes.

In this regard, retirement of the hardware or software may also be legally significant. If a system is no longer capable of lawful use, whether because of technical degradation, inability to maintain assurance evidence or inability to mitigate emerging legal risks, it should not remain in service. Article 36 review contributes to AI assurance not only by facilitating lawful use but by identifying when an AWS or AI-DSS should not be used.

Conclusion

The central challenge posed by AI-enabled military capabilities is not merely technical but socio-technical: lawfulness turns not on any single design feature but on the relationship between system function, system limitations, human control arrangements, operational concept and conditions of use. That relationship creates a translation problem that is relevant throughout the AI lifecycle. Legal requirements must be converted into technical specifications; technical evidence must be interpreted in legally meaningful terms; and military acquisition processes must be structured to bridge the gap between what legal advisers need to know and what development, testing and operational use actually reveal. Addressing that translation problem by embedding article 36 review across the lifecycle is the practical challenge this article has sought to address.

This article has argued that article 36 legal reviews should be understood as an integral component of the ADF's broader AI assurance framework, not as a narrow or final-stage compliance check. That interpretation is grounded both in the text of article 36 itself, which applies across the study, development, acquisition and adoption of a capability, and in the practical realities of AI-enabled systems that evolve through research, testing, integration and operational use. Consistent with the GC REAIM's *Responsible by Design* framework, SIPRI's analysis of responsible procurement, and Defence's Responsible AI Policy, meaningful legal review of military AI must be continuous, evidence based and structurally embedded.

A functional approach to legal reviews provides a practical methodology for making contextual, evidence-based determinations across the AI lifecycle. By focusing legal analysis on what a system does rather than how it is characterised, it enables a legal review to track the evolution of AI-enabled capabilities as they develop, are tested and are deployed.

The socio-technical dimension is central to legal assessment. As scholars have argued, the legal and ethical assessment of military AI cannot be resolved by identifying a human decision-maker nominally in the loop. It depends on the quality of the interaction between technical design, human cognition, interface design, institutional training and operational context. For article 36 review, this means that the key questions extend beyond what the system does in controlled conditions to whether the full human-machine system, in the circumstances of actual use, is capable of controlling its use in compliance with IHL. Where those conditions cannot be established or maintained, the system is not capable of lawful use regardless of its technical performance.

A lifecycle AI assurance model does not dilute technical innovation in the military. It makes it more consequential. By engaging legal reviews earlier, with more technical evidence and across more of the research, design and testing, it transforms article 36 from a late-stage compliance formality into a substantive governance instrument. By shaping design requirements, structuring procurement obligations and specifying training and operational conditions, article 36 review converts broad commitments to lawfulness, accountability and proportionate control into concrete assurance requirements that operate throughout the lifecycle of military AI. That is how article 36 can help give practical effect to the ADF's Responsible AI Policy and contribute to a broader assurance framework consistent with both IHL and the realities of contemporary military technology.

About the Author

Dr Damian Copeland is a distinguished international law practitioner and a leading authority on the legal review of new and emerging weapon systems, with particular expertise in autonomous weapon systems (AWS). He is the author of *A Functional Approach to the Legal Review of Autonomous Weapon Systems*, a foundational work in the field that bridges legal principles with technological realities. With decades of experience, Damian has advised on the legality of both conventional and advanced weapon technologies, offering strategic legal counsel to ensure compliance with international humanitarian law (IHL) and national legal frameworks. His work supports the lawful and responsible development of military capabilities, ensuring that legal considerations are integrated from the earliest stages of system design through to deployment. Damian is widely recognised for his contributions to advancing legal and ethical standards in the context of rapidly evolving defence technologies.

ENDNOTES

- 1 Australian Government, *2026 National Defence Strategy* (Canberra: Commonwealth of Australia, 2026), p. 76.
- 2 United Nations Institute for Disarmament Research, *The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain* (UNIDIR, 2026), pp. 7–19.
- 3 Department of Defence, *Policy Settings for Responsible Use of AI in Defence* (Canberra: Commonwealth of Australia, 2025).
- 4 J Dorsey and M Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human(e) Dimension', *International Law Studies* 106 (2025): 31.
- 5 Z Assaad and E Williams, 'Technology and Tactics: The Intersection of Safety, AI, and the Resort to Force', *Cambridge Forum on AI: Law and Governance* 1, e49 (2025): 7.
- 6 Ibid., p. 2.
- 7 Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM), *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (HCSS, 2025), pp. 13–18, 60–79.
- 8 N Goussac and V Boulanin, *Responsible Procurement of Military Artificial Intelligence* (SIPRI, 2026), pp. 3–6, 10–9.
- 9 D Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems* (Brill, 2024), pp. 1–20; GC REAIM, *Responsible by Design*, pp. 71–79.
- 10 *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)* (AP I), 1125 UNTS 3 (Diplomatic Conference on the Reaffirmation and Development of International Humanitarian Law Applicable in Armed Conflicts, 1977), article 36.
- 11 See Part III of Australian Defence Force, *Australia's Guide to the Legal Review of New Weapons, Means or Methods of Warfare* (Department of Defence, 2024).
- 12 GC REAIM, *Responsible by Design*, pp. 60–79.
- 13 Ibid., pp. 71–79.
- 14 Goussac and Boulanin, *Responsible Procurement of Military Artificial Intelligence*, p. vii.
- 15 UN Office for Disarmament Affairs, *Convention on Certain Conventional Weapons Group of Governmental Experts on Lethal Autonomous Weapons Systems: Revised Rolling Text as of 12 May 2025* (United Nations, 2025), pp. 1–8.
- 16 D Copeland, 'Article 36 Canaries in the Military AI Coalmine', *Article 36 Legal*, at: www.article36legal.com/blog/article-36-canaries-in-the-military-ai-coal-mine.
- 17 Department of Defence, *Policy Settings for Responsible Use of AI in Defence*.
- 18 *Protocol Additional to the Geneva Conventions of 12 August 1949*.
- 19 B Werner and B Shumeg, *Pillars of Assured and Trusted Artificial Intelligence Enabled Systems*, STO-MP-IST-210 (NATO Science and Technology Office, 2026).
- 20 Context-appropriate human control and judgement is currently subject to debate by states party to the Convention on Certain Conventional Weapons in the Group of Governmental Experts meetings on Lethal Autonomous Weapons Systems. UN Office for Disarmament Affairs, *Chair's Summary—First 2025 Session of the GGE on LAWS, CCW/GGE.1/2025/WP.1* (7 April 2025) paras 18–20, at: [https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_\(2025\)/CCW-GGE.1-2025-WP.1_-_Chair's_summary.pdf](https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2025)/CCW-GGE.1-2025-WP.1_-_Chair's_summary.pdf).
- 21 *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (OECD, 2019, amended 2024), at: <https://oecd.ai/en/ai-principles>.
- 22 Ibid.
- 23 Ibid.

- 24 'Australia's AI Ethics Principles', *Australian Government Department of Industry, Science and Resources* (website), at: www.industry.gov.au/publications/australias-ai-ethics-principles (accessed 12 May 2026).
- 25 'National Framework for the Assurance of Artificial Intelligence in Government', *Australian Government Department of Finance* (website), at: www.finance.gov.au/government/public-data/data-and-digital-ministers-meeting/national-framework-assurance-artificial-intelligence-government (accessed 12 May 2026).
- 26 Ibid.
- 27 Digital Transformation Agency, *Policy for the Responsible Use of AI in Government* (Commonwealth of Australia, 2025), at: www.digital.gov.au/ai/ai-in-government-policy.
- 28 Ibid., p. 5.
- 29 Digital Transformation Agency, *Technical Standard for Government's Use of Artificial Intelligence* (Commonwealth of Australia, 2025), at: www.digital.gov.au/policy/ai/AI-technical-standard.
- 30 'Essential AI Practices', *Australian Government National AI Centre* (website), at: www.ai.gov.au/staying-safe-and-responsible/essential-ai-practices.
- 31 GC REAIM, *Responsible by Design*.
- 32 Ibid., p. ix.
- 33 Ibid., p. ix.
- 34 Ibid., p. xi.
- 35 Third Summit on Responsible Artificial Intelligence in the Military Domain (REAIM), *REAIM Pathways to Action* (A Coruña: REAIM, 2026).
- 36 Ibid.
- 37 The 2026 REAIM *Pathways to Action* was signed by 36 states, in contrast to the 2024 REAIM *Blueprint for Action*, which was signed by over 60 states. See Z Assaad, 'Artificial Urgency: Reflecting on AI Hype at the 2026 REAIM Summit', *Just Security*, at: www.justsecurity.org/132504/ai-hype-2026-ream-summit.
- 38 Department of Defence, *Policy Settings for Responsible Use of AI in Defence*, p. iv.
- 39 Ibid., p. 5.
- 40 Ibid., p. 6.
- 41 There are presently 175 states party to Additional Protocol I. See ICRC *International Human Law Database*, at: <https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/state-parties>.
- 42 International Committee of the Red Cross (ICRC), *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare* (ICRC, 2006), pp. 9–18.
- 43 Ibid. See also Australian Defence Force, *Australia's Guide to the Legal Review of New Weapons*; and 'Legal Review Resources', *Article 36 Legal* (website), at: www.article36legal.com/legal-reviews-explained.
- 44 Ibid.
- 45 Based on the author's experience as the Director of Operations and International Law, responsible for conducting article 36 legal reviews on behalf of the Australian Government.
- 46 GC REAIM, *Responsible by Design*, pp. 60–79; Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 47–92.
- 47 Department of Defence, *Policy Settings for Responsible Use of AI in Defence*, p. 6.
- 48 Ibid., p. 5.
- 49 GC REAIM, *Responsible by Design*.
- 50 ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, pp. 17–18.
- 51 *Protocol Additional to the Geneva Conventions of 12 August 1949*.
- 52 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 93–146; United Nations Institute for Disarmament Research, *The Global Prism of Military AI Governance*, pp. 13–9.
- 53 GC REAIM, *Responsible by Design*, p. 79.
- 54 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 21–46.
- 55 *Geneva Convention (III) Relative to the Treatment of Prisoners of War*, 75 UNTS 135 (Geneva, 12 August 1949), article 5.

- 56 Australian Government National AI Centre, 'Guidance for AI Adoption Foundations', *Australian Government National AI Centre* (website), at: <https://www.ai.gov.au/staying-safe-and-responsible/essential-ai-practices/guidance-ai-adoption-foundations> (accessed 25 June 2026): (1) Decide who is accountable, (2) Understand impacts and plan accordingly, (3) Measure and manage risks, (4) Share essential information, (5) Test and monitor, (6) Maintain human control.
- 57 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, Chapter 7.
- 58 ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, p. 9.
- 59 Werner and Shumeg, *Pillars of Assured and Trusted Artificial Intelligence Enabled Systems*.
- 60 GC REAIM, *Responsible by Design*, p. 71.
- 61 *Protocol Additional to the Geneva Conventions of 12 August 1949*, article 57.
- 62 *Ibid.*, article 48.
- 63 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 93–94.
- 64 *Ibid.*, p. 174.
- 65 *Ibid.*, p. 174.
- 66 GC REAIM, *Responsible by Design*, p. 9.
- 67 *Ibid.*, p. 9.
- 68 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, Chapter 8, refers to the 'informative phase'.
- 69 See for example 'Article 36 Legal Compliance Reports', *Article 36 Legal*, at: www.article36legal.com/compliance-reports.
- 70 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 199–200.
- 71 GC REAIM, *Responsible by Design*, p. 43.
- 72 *Ibid.*, p. 9.
- 73 *Ibid.*, pp. 26–33.
- 74 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 217–219.
- 75 *Ibid.*, pp. 183–184.
- 76 ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, p. 14; Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, p. 152.
- 77 'Lawful by Design Initiative', *Article 36 Legal*, at: <https://www.article36legal.com/lawful-by-design>.
- 78 Goussac and Boulain, *Responsible Procurement of Military Artificial Intelligence*, pp. 5–6.
- 79 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 138–139.
- 80 *Ibid.*, pp. 129.
- 81 *Ibid.*, p. 233.
- 82 GC REAIM, *Responsible by Design*, p. 9.
- 83 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 197–199.
- 84 *Ibid.*, p. 198.
- 85 *Ibid.*, p. 32.
- 86 ICRC, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare*, p. 14.
- 87 GC REAIM, *Responsible by Design*, p. 10.
- 88 *Ibid.*, pp. 200–204.
- 89 GC REAIM, *Responsible by Design*, p. 10.
- 90 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 149–152.
- 91 The AutoPractices Project, *Strengthening Human Agency in the Military Domain: Best Practices Toolkit for Policymakers, Developers, and Users of AI Systems* (Odense: Center for War Studies, University of Southern Denmark, 2026), pp.19–20.
- 92 Z Assaad and E Williams, 'Technology and Tactics: The Intersection of Safety, AI, and the Resort to Force', *Cambridge Forum on AI: Law and Governance* 1, e49 (2026): 6–11.

- 93 K Klonowska and T Woodcock, *Rhetoric and Regulation: The (Limits of) Human/AI Comparison in Legal Debates on Military AI*, ASSER Research Paper No. 2025-07 (T.M.C. Asser, 2025).
- 94 GC REAIM, *Responsible by Design*, p. 23 (n. 5).
- 95 For example, UK Ministry of Defence, *JSP 936 V1.1: Dependable Artificial Intelligence (AI) in Defence* (UK Ministry of Defence, 2024), p. i, at: https://assets.publishing.service.gov.uk/media/6735fc89f6920bf-b5abc7b62/JSP936_Part1.pdf.
- 96 D Copeland, R Liivoja and L Sanders, 'The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems', *Journal of Conflict and Security Law* 28, no. 2 (2023): 285–316, 300.
- 97 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, p. 227.
- 98 GC REAIM, *Responsible by Design*, p. 47.
- 99 N Jevglevskaia, *International Law and Weapons Review* (Cambridge University Press, 2022), p. 195.
- 100 Copeland et al., 'The Utility of Weapons Reviews', pp. 300–304.
- 101 Copeland, *A Functional Approach to the Legal Review of Autonomous Weapons Systems*, pp. 237–239.
- 102 GC REAIM, *Responsible by Design*, p. 51.
- 103 Ibid., p. 239.

/ TEVV AND THE RESPONSIBLE AND LAWFUL USE OF MILITARY AI SYSTEMS

Manoj Harjani

Introduction

Reportage and analysis of ongoing conflicts in Ukraine, Gaza and, more recently, Iran are highlighting the growing use of artificial intelligence (AI) in the military domain.¹ In Ukraine, the spotlight has been on how AI has enabled targeting for drones, which have dominated the battlefield there. A case in point is the TFL-1, a module developed by the Ukrainian company The Fourth Law, which leverages machine vision to enable drones to autonomously strike a target that has been identified by an operator.² Moving to Gaza, attention has focused on Israel's use of AI-enabled decision-support systems to identify targets. The Lavender system developed by the Israel Defense Forces (IDF) reportedly generated a list of 37,000 human targets suspected to be members of Hamas and Palestinian Islamic Jihad, using a wide variety of sources.³ The IDF allegedly utilised this AI-generated list of targets during the early stages of its military response to the October 2023 attack by Hamas, conducting air strikes supported by other AI-enabled systems.⁴ Turning to Iran, where a missile strike by the United States in February 2026 on a girls' school is estimated to have killed up to 168 people,⁵ there has been considerable debate around the role of AI in the target selection process.⁶ This debate is complicated by confusion over the facts—discussion zeroed in on the use of Claude, a generative AI-enabled chatbot developed by the American company Anthropic, when the targeting for the US campaign against Iran was primarily supported by the Maven system, which leverages machine learning to analyse satellite imagery and sensor data to identify targets.⁷

Meanwhile, multilateral dialogue aimed at creating regulatory guardrails for the use of AI in the military domain has achieved relatively little so far and is moving at a much slower pace in comparison. For instance, at the third Responsible AI in the Military Domain (REAIM) Summit, held in A Coruña, Spain in February 2026, only 44 out of 85 states that attended endorsed the outcome document, 'Pathways to Action'.⁸ Compared to the outcome documents from the first and second REAIM summits, held in 2023 and 2024, which garnered the endorsement of around 50 and 60 states respectively, the drop in

support for the latest outcome document also sent a worrying signal regarding a potential loss of interest among states in working to develop regulatory guardrails for the use of AI in the military domain.⁹ Moreover, this is the first REAIM outcome document to not have the endorsement of at least one of the two superpowers.¹⁰ China and the United States both endorsed the ‘Call to Action’¹¹ outcome document from the inaugural REAIM summit, held in 2023, while only the United States supported the ‘Blueprint for Action’¹² outcome document from the second REAIM summit, in 2024. In a similar vein, the US-led Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (US Political Declaration),¹³ launched in 2023—which garnered the support of nearly 60 states¹⁴—has been quietly set aside by the Trump administration since it took office in 2025.

Taken together, it is clear that facts on the ground are evolving much more quickly and at a scale that multilateral dialogue cannot keep up with. In such circumstances, an important question that arises is around what can be done to ensure the responsible and lawful use of military AI systems. A cynical answer to this question would point to the fact that states will do what they feel is necessary since they are effectively unconstrained in leveraging AI to apply or guide the use of force in the pursuit of political and strategic objectives. Examples from Ukraine, Gaza and Iran appear to validate this point of view, suggesting that states are not particularly interested in the responsible and lawful use of military AI systems. However, this cynical reading also obscures important facts. First, military AI systems do not operate in a vacuum—they are ‘complex sociotechnical systems ... [consisting] of material and social components which, by being put into particular kinds of relations, work together in specific ways’.¹⁵ Second, militaries have to reckon with the limitations of AI systems in terms of their predictability and understandability—what is colloquially referred to as the ‘black box dilemma’.¹⁶ Finally, there is a tension between trust and control when it comes to military AI systems, and states face a number of challenges in aiming to reconcile these desired outcomes.¹⁷

At the same time, even if multilateral dialogue on regulatory guardrails for military AI systems can reach a consensus among states, the way in which its output is framed leaves considerable room for interpretation and significant potential for gaps in accountability. For example, the REAIM Blueprint for Action document makes reference to military AI systems being ‘developed, deployed and used in accordance with international law’¹⁸ but does not delve into how this might be achieved. Additionally, the Blueprint for Action document points to the necessity of military AI systems being ‘ethical and human-centric’¹⁹ but similarly does not outline how this can be concretely implemented in practice. While the REAIM Pathways to Action document sought to provide more detail—for example, by highlighting the importance of a ‘responsible by design’ approach across the lifecycle of military AI systems²⁰—it is also framed in largely abstract and aspirational terms. Nevertheless, the Pathways to Action document outlines several recommendations to operationalise the various principles identified across REAIM outcome documents at both the national and international levels. Among these recommendations is the

promotion of ‘robust testing, evaluation, validation and verification (TEVV) and [the integration of] TEVV requirements in relevant policies, doctrines, and procurement processes’.²¹ TEVV was also highlighted in the US Political Declaration in the context of ensuring ‘that the safety, security, and effectiveness of military AI capabilities are subject to appropriate and rigorous testing and assurance within their well-defined uses and across their entire life-cycles’.²²

This article seeks to explore the role of TEVV in supporting emerging regulatory guardrails that are aiming to ensure the responsible and lawful use of military AI systems. The first and arguably most critical contention behind this is that TEVV is an essential enabler for governance and regulation of military AI systems, as it could potentially provide a way to translate abstract and aspirational concepts into tangible practices that states can undertake. A second and equally important contention is that TEVV is not a silver bullet when it comes to regulatory guardrails, and comes with its own set of challenges that states must find ways to navigate. Three sets of challenges associated with implementing TEVV to support regulatory guardrails are identified and discussed: (1) the lack of standardised TEVV practices for military AI systems, (2) variations in implementing TEVV, and (3) the private sector’s involvement. A mixture of qualitative methods—documentary analysis and close reading of primary and secondary sources—has been employed alongside a small-N case study design for comparing approaches to TEVV for military AI systems in Australia, the United Kingdom and the United States. The structure of the article is as follows: an overview of TEVV, a discussion of how TEVV can support regulatory guardrails for the responsible and lawful use of military AI systems, a comparison of state-level approaches across Australia, the United Kingdom and the United States, and a discussion of key challenges associated with leveraging TEVV to support regulatory guardrails for military AI systems.

What is TEVV?

‘Testing, evaluation, verification and validation’ (TEVV) refers to a set of practices without a universal definition or set of standards in the military domain.²³ The concept itself and its four elements are not used consistently—testing and evaluation is often treated separately from verification and validation, and the order of the latter two elements is sometimes swapped. Additionally, a fifth element is sometimes associated with verification and validation—namely, accreditation. The various elements within TEVV—regardless of whether they are treated separately as T&E and V&V (or VV&A)—share a common purpose, which is risk reduction.²⁴ They are also all grounded in the broader set of practices associated with assurance, which for AI could cover a wide range of possible qualities associated with a system’s outcomes, including that they are ‘valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users’.²⁵ The ways in which desired outcomes for assurance are identified can vary considerably—this is both essential

and problematic when it comes to how TEVV can support regulatory guardrails for the responsible and lawful use of military AI systems. For the purpose of this article, the table below outlines one possible way to define the four elements within TEVV (and assurance), which was chosen for its simplicity and succinctness (see Table 1).

Table 1—Definitions of elements within TEVV and of assurance

Testing	Using pass/fail criteria to check a specific, single function. Testing is usually associated with verification
Evaluation	Combining multiple tests and experimental use to provide an overall judgement. Evaluation typically relates to validation
Verification	Checking against a specification. It asks, ‘Did we build the thing right?’ or ‘Are we doing the thing right?’
Validation	Checking against an end user’s need. It asks, ‘Did we build the right thing?’ or ‘Are we doing the right thing?’
Assurance	Providing a compelling case for the trustworthiness of a system

Note. Adapted from Defence Science and Technology Laboratory, *Assurance of Artificial Intelligence and Autonomous Systems: A Dstl Biscuit Book* (London: Ministry of Defence, 2021), pp. 4–5.

Nevertheless, there are other ways of contextualising TEVV. For example, in the civilian domain, TEVV can be framed as one of several paradigms within the practice of AI evaluation, which can be described as a ‘process of measuring and anticipating the behavioural properties of AI systems and their societal impact to inform decisions about their use’.²⁶ Other paradigms within AI evaluation include benchmarking²⁷ and measuring real-world impact.²⁸ Within AI evaluation, each paradigm has a distinct focus, and for TEVV this is ‘ensuring that AI systems behave in a well-defined and predictable way’.²⁹ AI evaluation is closely linked to the field of AI safety, which has largely been concerned with the existential risks associated with advanced AI systems, although this has created its own challenges in terms of directing attention away from existing work—including through TEVV—on ensuring the robustness and reliability of AI systems.³⁰ Multilateral dialogue on AI safety has also evolved considerably since the inaugural AI Safety Summit organised by the United Kingdom in 2023. Twenty-nine states—including China and the United States—signed the Bletchley Declaration that was produced through that meeting.³¹ In 2024 at the AI Seoul Summit, the framing of the outcome document shifted to encompass not just safety but also innovativeness and inclusion.³² Similarly, the 2025 AI Action Summit held in Paris changed the focus of the outcome document again to include mention of inclusivity and sustainability.³³ At the 2026 AI Impact Summit held in New Delhi, the scope of the outcome document broadened even further, covering themes ranging from democratising AI to resilience, innovation and efficiency.³⁴ Little to no attention has been paid to TEVV in the outcome documents produced since the first AI Safety Summit in 2023.

TEVV for Military AI Systems

TEVV for military AI systems³⁵ does not take place in a vacuum and is shaped by existing practices for other types of systems.³⁶ This is likely to include the general set of TEVV practices associated with all systems developed or purchased by a military, as well as specific practices depending on the nature of the system.³⁷ For example, if the system is a helicopter, there will be TEVV practices from the army aviation sector to consider. Similarly, each service branch of a military may have TEVV practices tailored to its needs and influenced by its organisational, engineering and safety cultures. For militaries in states that may not have the capacity or resources to maintain an in-house capability, their TEVV practices will depend heavily on the supplier they are purchasing a system from. Moreover, when it comes to AI systems, there is a need to consider the TEVV practices of the private sector companies developing them, since they are at the forefront of innovation in this field. For the most advanced AI systems, militaries will be dealing with the handful of companies that have the necessary capabilities. All of these considerations point to TEVV for military AI systems being a complex and costly endeavour. This in turn raises the question of whether this ‘premium’ associated with fielding military AI systems will create a clear gap in terms of their impact on the balance of power, which challenges the conventional wisdom suggesting that access to AI is relatively democratised and that more states will have the ability to leverage AI in the military domain.³⁸

From a technical perspective, TEVV for military AI systems must also account for the distinctive features of AI-enabled systems more generally. This entails paying attention to (1) the need for continuous testing and monitoring due to the dynamic nature of AI-enabled systems, particularly in terms of their ability to learn and adapt, (2) the need to evaluate unexpected and unanticipated system behaviours arising from the ‘black box dilemma’, and (3) the need to assess the security and resilience of AI-enabled systems given their vulnerability to adversarial attacks.³⁹ Regarding continuous testing and monitoring, this necessitates a different approach to TEVV that moves away from treating it as a discrete stage prior to fielding a system. It has also prompted exploration of potential ways to automate TEVV.⁴⁰ To address the ‘black box dilemma’, one possible way forward is through implementing safety middleware that can ‘supersede the true system’s decisions under certain explicit, well-defined conditions’,⁴¹ thereby making it easier to verify specific conditions under which certain behaviours of the system are occurring. Finally, when it comes to adversarial attacks, these introduce a significant degree of complexity, as there are a wide range of possible attacks, from ‘poisoning’ of training data to the ‘extraction’ of a model through reverse engineering of its output.⁴²

TEVV and regulatory guardrails

Turning to the role that TEVV plays in supporting regulatory guardrails for the responsible and lawful use of AI in the military domain, there are three aspects to unpack further: (1) providing the means to tangibly measure desired outcomes of an AI system, (2) enhancing human agency across the lifecycle of AI systems, and (3) enabling the conduct of legal reviews. These three aspects of TEVV's role were selected to highlight that it is more than a set of practices to narrowly certify a system as ready for use or fit for purpose.

Measuring Desired Outcomes

At the heart of responsible and lawful use of AI in the military domain are the desired outcomes of trust and control.⁴³ These are typically framed as expectations of military AI systems and of the relationship between these systems and their operators, which can occur in the context of human–machine teams.⁴⁴ Put simply, for a military to use an AI system responsibly and lawfully, there must be trust in that system and its operator, but also between the operator and the system. Trust in a military AI system also depends in part on an operator's ability to control that system to function in specific ways. The ability of a military to meet legal obligations arising from using an AI system also depends on control over that system, particularly through rules that circumscribe the behaviour of the system and its operator. Trust therefore implies 'a "contract" between humans and AI',⁴⁵ whereas control focuses on 'the capacity of humans to ensure, in a rather deterministic way, specific (or, at least, sufficiently tightly constrained) behaviours of AI-driven systems'.⁴⁶ Reconciling trust and control becomes challenging depending on the circumstances related to a specific application of a military AI system. This is where TEVV can play an important role, particularly in measuring the desired outcomes of trust and control. For example, TEVV can be used to measure trust based on reliability or predictability⁴⁷ and to measure control by providing evidence of an AI system's patterns of behaviour through testing. It can also measure trust based on values, preferences and beliefs⁴⁸ where there has been an appropriate translation into expected behaviours and output of an AI system that can be evaluated.

However, TEVV is not a silver bullet. Measuring trust and control in military AI systems faces a number of interrelated challenges. Six of these are identified here, with the caveat that this is not an exhaustive list.⁴⁹ First, the appropriate unit of analysis for TEVV in military AI systems can be difficult to define, particularly when the same AI model is used for multiple applications, when it is part of a larger system, and when it is part of a human–machine team.⁵⁰ Second, TEVV depends on:

clearly stating [a] purpose and being able to assess whether the tool serves it as well as potential alternatives, and whether 'ground truth' benchmarks for evaluating it accurately reflect contexts in which [it] will be used.⁵¹

But for military AI systems, ‘ground truth’ benchmarks are difficult to define given their unpredictable and inscrutable nature. Moreover, identifying the different contexts in which military AI systems will be used is not a straightforward task—modelling and simulation methods and techniques that are traditionally employed in TEVV may not be sufficient. Third, and related to the previous challenge, is the fact that TEVV will not be able to anticipate all the conditions and environments under which military AI systems will operate, creating a potentially impractical number of scenarios to test and evaluate against.⁵² Fourth, military AI systems share the same challenges associated with all AI systems in terms of difficulty in predicting and understanding how they might fail, which complicates the scope of TEVV.⁵³ Fifth, although military AI systems are likely to comprise software components that were developed by the private sector, the applicability of commercial TEVV tools is not straightforward for a military context.⁵⁴ Finally, military AI systems which have models that are updated through data gathered during their use necessitate some parts of the TEVV process to be carried out continuously.⁵⁵

Overcoming these challenges may require militaries to rethink the practice of TEVV from a socio-technical perspective in order to:

enable a broader understanding of AI impacts and the key decisions that happen throughout, and beyond, the AI lifecycle—such as whether technology is even a solution to a given task or problem.⁵⁶

TEVV that leverages a socio-technical perspective could potentially incorporate practices from a wider range of fields in order to provide novel or more effective ways to measure trust and control in military AI systems. However, several of the challenges identified also underscore the point made earlier that TEVV for military AI systems is a complex and costly endeavour. Along with potentially allocating dedicated resources catering to the unique demands of TEVV for military AI systems, militaries are likely to have to invest in developing new TEVV practices and finding ways to adapt existing practices. However, this poses an additional challenge of its own, given the pressure created by militaries’ increasing pace of AI adoption. It will not be straightforward to update existing TEVV practices or introduce new ones under these circumstances where there is an imperative to deploy systems more quickly.

Enhancing Human Agency

The importance of human agency for the responsible and lawful use of military AI systems across all service branches of a military cannot be overstated. In an army context, for example, human agency takes on particular significance given close contact with civilian populations. The REAIM Call to Action document refers to how AI can ‘shape and impact decision making, and [the importance of ensuring] that humans remain responsible and accountable for decisions when using AI in the military domain’.⁵⁷

It also specifically mentions the ‘importance of ensuring appropriate safeguards and human oversight of the use of AI systems, bearing in mind human limitations due to constraints in time and capacities’.⁵⁸ The REAIM Blueprint for Action document similarly refers to how:

Appropriate human involvement needs to be maintained in the development, deployment and use of AI in the military domain, including appropriate measures that relate to human judgement and control over the use of force.⁵⁹

Meanwhile, the REAIM Pathways to Action document makes specific reference to how the:

nature and degree of human involvement should be appropriate considering, among other factors, the operational context, the function performed, the technical characteristics and capabilities, as well as human factors such as training and fatigue, and the risks and benefits involved.⁶⁰

In the US Political Declaration, there is mention of ‘senior officials effectively and appropriately oversee[ing] the development and deployment of military AI capabilities with high-consequence applications, including, but not limited to, such weapon systems’⁶¹ as well as ensuring that ‘relevant personnel exercise appropriate care in the development, deployment, and use of military AI capabilities, including weapon systems incorporating such capabilities’.⁶² What is missing in all of these references to human agency is how it can be implemented in practice.

A gradual evolution in how human agency is being understood—particularly in terms of going beyond a ‘narrow focus on human operators executing military missions at the final tactical level’⁶³—has paved the way for its consideration across the entire lifecycle of a military AI system. In the nine-stage lifecycle framework developed by the IEEE Standards Association, TEVV is both a distinct stage and an ongoing activity where human decision-makers play a key role (see Figure 1). As a stage of the lifecycle, TEVV is situated between ‘procurement and acquisition’, and ‘considering the human: education, training, and human-system integration.’ This position reflects TEVV’s role as a bridge between dimensions of an AI system that are technical in nature and those that are dependent on a human user’s needs and interaction with the system. As an ongoing activity taking place throughout each stage of the lifecycle, TEVV is considered together with monitoring, hardware or software updates, interoperability and maintenance, which leans more towards the technical and operational dimensions of AI systems. Specific best practices related to TEVV that can enhance human agency have been identified in a study conducted by the University of Southern Denmark. They include testing the interface of military AI systems not only from a technical and functional perspective but also in terms of human users’ understanding and interaction, verifying that intended users

of military AI systems are empowered to use them in ways that are legally and ethically compliant, and engaging in TEVV to independently verify commercial vendors' claims about their military AI systems.⁶⁴

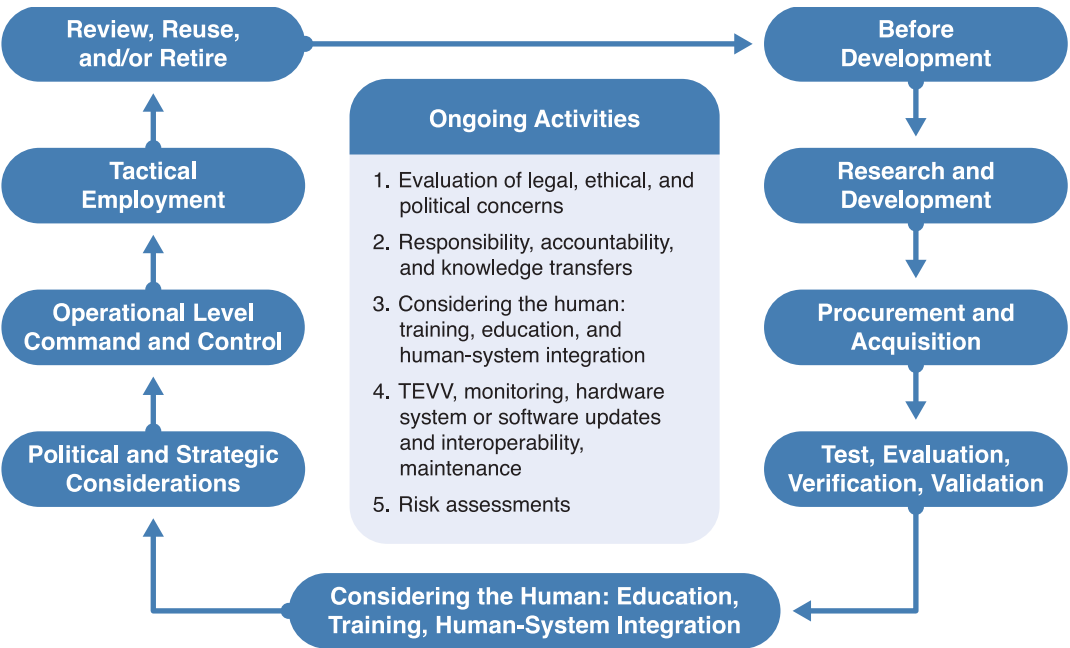


Figure 1—Lifecycle framework for military AI systems

Note. Adapted from IEEE Standards Association Industry Connections Research Group on Issues of Autonomy and AI for Defence Systems, *A Framework for Human Decision-Making Through the Lifecycle of Autonomous and Intelligent Systems in Defence Applications* (New York NY: IEEE Standards Association, 2024), p. 21, at: <https://ieeexplore.ieee.org/document/10707139>.

Enabling Legal Reviews

Both the REAIM Blueprint for Action and Pathways to Action documents make specific references to legal reviews as a regulatory guardrail for military AI systems. In the Blueprint for Action, legal reviews are highlighted alongside trust and confidence-building measures as well as risk-reduction measures,⁶⁵ whereas the Pathways to Action document identifies the conduct of legal reviews as a measure to operationalise REAIM principles at the national level.⁶⁶ In the US Political Declaration, legal reviews are specifically mentioned as a step that states can take to ‘ensure that their military AI capabilities will be used consistent with their respective obligations under international law, in particular international humanitarian law’.⁶⁷ While the exact way in which legal

reviews are carried out varies considerably, they are generally understood as an obligation under article 36 of Additional Protocol I to the Geneva Conventions.⁶⁸ The starting point for legal reviews for military AI systems is the consensus that has emerged around the applicability of international humanitarian law (IHL), but the challenge has been in determining exactly how it applies. This challenge is exacerbated by a few factors, such as (1) the wide range of technologies embedded within military AI systems, which may be part of other, larger systems involving consideration of other technologies, (2) whether AI systems are fundamentally incompatible with IHL principles due to their unpredictable and inscrutable nature, and (3) the dynamics created through human–machine teams and human–machine interaction.⁶⁹

TEVV has the potential to address some of these factors when integrated within the conduct of legal reviews. This would bring technical and legal assessments of military AI systems together, with significant implications for current expertise, methods and processes related to both TEVV and legal reviews. For instance, IHL principles could be translated into technical criteria that are then ‘embedded into the system during the designing phase. Once the legal standards are learned by the system, compliance with those standards is a technical task’.⁷⁰ This approach is aligned with what has been proposed by the Global Commission on Responsible AI in the Military Domain (GC REAIM), which has been encapsulated in the concept of ‘responsible by design’.⁷¹ Although one of the GC REAIM’s core recommendations is that TEVV should inform national policies guaranteeing human responsibility across a military AI system’s lifecycle,⁷² the link between TEVV and legal reviews is not made explicit, and they are treated as separate activities and sets of practices. However, the fact that experts and practitioners in both TEVV and legal reviews are increasingly pointing to the need for the two activities and sets of practices to evolve to better address the challenges posed by military AI systems is a good sign. For example, recognising the limitations of conventional legal reviews—particularly when they are conducted as a ‘discrete, one-time requirement performed during the weapon acquisition phase, typically just prior to adoption’⁷³—will pave the way for closer alignment between TEVV and legal review processes, which need to be conducted throughout the lifecycle of military AI systems. At the same time, there are some complications that can arise from integrating TEVV and legal reviews. The view of the International Committee of the Red Cross, which is entrusted by the international community of states as the guardian of IHL,⁷⁴ is that:

it is essential to preserve human control over tasks and human judgement in decisions that may have serious consequences for people’s lives in armed conflict, especially where they pose risks to life, and where the tasks or decisions are governed by specific rules of international humanitarian law.⁷⁵

Integrating TEVV and legal reviews could therefore run the risk of transferring responsibility for legal determinations away from humans.

Comparing state-level approaches

Australia, the United Kingdom and the United States represent a spectrum of different types of actors in the ongoing multilateral dialogue aimed at creating regulatory guardrails for the use of AI in the military domain. The United States sits at one end of the spectrum in terms of the size and capabilities of its military. As treaty allies of the United States, Australia and the United Kingdom not only draw on the capabilities of the United States but also contribute to joint capabilities through arrangements such as AUKUS. Through AUKUS, the three states are also pursuing extensive integration of their defence industrial and technological bases.⁷⁶ The United Kingdom and United States are also both permanent members of the United Nations Security Council and founding members of the North Atlantic Treaty Organisation. Australia and the United States are part of the Quad, and all three states are also part of an intelligence-sharing arrangement known as the Five Eyes. The relative postures and positioning of the three states when it comes to AI in the military domain is not a straightforward comparison, but the table below outlines some indicators with respect to regulatory guardrails (see Table 2). These indicators should be viewed with a measured eye—publicly articulated statements in a defence AI policy or strategy and the endorsement of non-binding statements such as the REAIM outcome documents and the US Political Declaration cannot be construed as evidence of a state’s actual position in practice.

Table 2—Indicators for posture and positioning on regulatory guardrails for AI in the military domain

	Defence AI Policy / Strategy	REAIM Call to Action (2023)	REAIM Blueprint for Action (2024)	REAIM Pathways to Action (2026)	US Political Declaration (2023)
Australia	2026	✓	✓	✓	✓
United Kingdom	2022	✓	✓	✓	✓
United States	2026	✓	✓	✗	✓

Compiled by the author.

Australia

In March 2026, Australia’s Department of Defence (Defence) released a document titled *Policy Settings for Responsible Use of Artificial Intelligence in Defence* (Policy Settings),⁷⁷ which outlined its approach to responsible AI in the military domain across the lifecycle of military AI systems. The Policy Settings document is explicitly framed as a follow-up to implement Australia’s commitments made through its endorsement of the REAIM Call to Action and Blueprint for Action documents, as well as the US Political

Declaration.⁷⁸ It anchors Australia's obligations for responsible use of AI in the military domain on three policy requirements: lawfulness, adherence to values-based principles, and proportionate controls.⁷⁹ Lawfulness is based on compliance with domestic law and international legal obligations, clear accountability through identifiable officers, and the conduct of legal reviews.⁸⁰ Under adherence to values-based principles, five principles have been identified: accountable, bias and harm mitigation, explainable, human-centric, and reliable and secure.⁸¹ These principles have been developed in line with Australia's 2024 Policy for the Responsible Use of AI in Government⁸² and 2019 AI Ethics Principles.⁸³ Finally, proportionate controls comprise risk-based 'layers of policies, processes, training, and procedures, including ongoing assurance and after-action evaluation'.⁸⁴ TEVV is included as a control measure, alongside post-incident/impact reviews, and relevant reporting and oversight mechanisms.⁸⁵ However, the role of TEVV is otherwise not elaborated in the Policy Settings document.

Defence's 2021 Test and Evaluation Strategy⁸⁶ only mentions AI once in the context of emerging technologies necessitating new approaches to TEVV.⁸⁷ This is due to the fact that it is meant to cover all systems regardless of the technologies they comprise. In its 2024 Defence Industry Development Strategy, Defence identified TEVV as a 'sovereign defence industrial priority', with specific mention of building skill level and capacity in TEVV across warfighting domains with respect to AI.⁸⁸ This was an important signal, both in terms of Defence's overall assessment of current TEVV capabilities and in terms of its expectations of how Australia's defence TEVV workforce, processes and practices should evolve. It also built on the 2020 Concept for Robotic and Autonomous Systems,⁸⁹ where TEVV is explicitly linked to determining trust in military AI systems.⁹⁰ In this document, TEVV also supports the traceability of decisions made by military AI systems, which in turn contributes to trust.⁹¹ Additionally, reform of existing TEVV processes and practices 'to understand how to verify systems that are the subject of continual improvement while balancing this need against the necessity of providing capability certainty'⁹² is highlighted. Moreover, with the Defence AI Centre that was set up in 2024, there is now dedicated bureaucratic capacity for coordination that will also shape the evolution of TEVV, including through initiatives such as developing common standards for military AI systems.⁹³

Turning to service-level approaches, in the second version of its Robotic and Autonomous Systems Strategy,⁹⁴ the Australian Army mentions TEVV in two lines of effort: first, as part of developing, experimenting and prototyping military AI systems;⁹⁵ and second, as part of exploring and integrating military AI systems within force design, specifically in terms of providing assurance for 'fit-for-purpose capabilities into service at the right time and quantities'.⁹⁶ For the first line of effort, TEVV is intended to 'test and demonstrate key capabilities in representative environments and is critical to ensuring system reliability, robustness and usability in contested environments and to inform Army's user needs',⁹⁷ whereas the second line of effort sees TEVV supporting future force design.⁹⁸ In the

Royal Australian Navy's 2020 RAS-AI Strategy 2040,⁹⁹ the need for qualified TEVV practitioners is highlighted,¹⁰⁰ as well as the role that TEVV plays in providing 'proof that risk is contained within acceptable boundaries'.¹⁰¹ The Navy's strategy also outlines how AI can be an opportunity for TEVV in terms of automating testing, documentation, data processing, decision-making and system control, among other areas.¹⁰² While the Royal Australian Air Force does not have a comparable strategy document related to military AI systems, its efforts under the 2015 Plan Jericho¹⁰³ led to a dedicated Test and Evaluation Directorate being set up within the Air Warfare Centre inaugurated in 2016.¹⁰⁴

Australia's approach to TEVV for military AI systems is generally headed in a positive direction. TEVV has been clearly prioritised at the strategic level centrally in the Australian Defence Force (ADF) through documents such as the 2024 Defence Industry Development Strategy and 2020 Concept for Robotic and Autonomous Systems. Service-level strategies and initiatives for military AI systems also point to TEVV being recognised as a priority, including in the Army's 2022 Robotic and Autonomous Systems Strategy, the Navy's RAS-AI Strategy 2040, and the Air Force's 2015 Plan Jericho. However, the actual impact of this strategic prioritisation on the ADF's TEVV capabilities remains to be seen. Each service branch has high-profile initiatives involving military AI systems—the Army's M113 AS4 optionally crewed combat vehicle,¹⁰⁵ the Navy's Bluebottle uncrewed surface vessel,¹⁰⁶ and the Air Force's MQ-28A Ghost Bat uncrewed combat aerial vehicle supporting crewed aircraft¹⁰⁷—which means that their TEVV capabilities, processes and practices will be evolving through experience on the fly. Moreover, collaboration through AUKUS is also likely to shape Australia's defence TEVV enterprise. An example is the trialling of AI-enabled decision-support systems onboard the Air Force's P-8A Poseidon aircraft during Exercise Talisman Sabre held from July to August 2025.¹⁰⁸ Although AUKUS has faced some uncertainties,¹⁰⁹ it remains an important platform for TEVV-related activities for military AI given its emphasis on experimentation and technology development, particularly under Pillar II.

United Kingdom

In 2022, the United Kingdom's Ministry of Defence (MoD) outlined its approach for the delivery of AI-enabled defence capabilities based on three guiding principles: (1) ambition in terms of tools and operational effects, (2) safety, and (3) responsibility.¹¹⁰ In the same document it also identified five ethical principles for military AI systems: (1) human-centricity, (2) responsibility, (3) understanding, (4) bias and harm mitigation, and (5) reliability.¹¹¹ Building on this approach, the 2022 Defence Artificial Intelligence Strategy¹¹² articulated four objectives: (1) transforming into an 'AI ready' organisation, (2) adopting and exploiting AI at pace and scale, (3) strengthening the domestic defence and security AI ecosystem, and (4) shaping global developments to promote security, stability and

democratic values.¹¹³ Furthermore, in 2024, the MoD issued the first part of a joint service publication, *JSP 936 V1.1: Dependable Artificial Intelligence (AI) in Defence*,¹¹⁴ which contained a directive describing the mandated direction related to the responsible and lawful use of military AI systems. The second part of JSP 936, which contains a defence AI ethical risk assessment toolkit, has not been made public in its entirety. JSP 936 builds on a family of related MoD policies, including on the management of health and safety (JSP 375¹¹⁵), acquisition safety (JSP 376¹¹⁶), information, knowledge, digital and data (JSP 441117), human factors integration (JSP 912¹¹⁸), and modelling and simulation (JSP 939¹¹⁹).

In January 2026, the MoD issued updated guidance on how its approach to TEVV is changing in a document titled *Test and Evaluation (T&E): Future Advantage Through Evaluation (FATE)*.¹²⁰ The challenge posed by AI for TEVV is highlighted in this document, along with AI's potential to transform TEVV. Crucially, the document also highlights the need for a different approach to TEVV and a slew of initiatives to support the United Kingdom's TEVV enterprise, including a dedicated innovation fund, increased support for training, and better coordination with industry. The 2022 Defence Artificial Intelligence Strategy outlines specific measures related to TEVV for military AI systems, including new live and virtual test capabilities, establishing a framework for TEVV of military AI systems across their lifecycle that includes both the technical and human dimensions, and driving new technical standards and regulations.¹²¹ JSP 936 also provides detailed guidance on TEVV in the context of AI assurance,¹²² which has been complemented by an AI Assurance Framework developed by the MoD's Defence AI Centre that covers who is responsible, the relevant steps involved and the required documentation for evidence, and identifies enabling technologies.¹²³ The Defence AI Centre has also published an *AI Practitioner's Handbook*¹²⁴ with a corresponding resource library containing templates and tools to support TEVV.¹²⁵ At the service level, the British Army and Royal Navy have so far published strategies for AI, but only the Army's 2023 document is available publicly¹²⁶ and there is limited mention of TEVV's role within it.

The United Kingdom's approach to TEVV for military AI systems appears well organised, with clear prioritisation by the MoD. A review by the House of Commons Defence Committee published in January 2025 did not make any recommendations directly mentioning TEVV, but devoted a section to addressing challenges posed by AI for defence procurement, such as shorter development cycles, evolving contractor–supplier relationships, and the role of primes versus smaller companies.¹²⁷ Any changes to address these challenges will also affect the TEVV enterprise. Nevertheless, with JSP 936 and the Defence AI Centre, the MoD is well positioned to drive the evolution of the United Kingdom's defence TEVV to meet the challenges and needs of military AI systems and to ensure responsible and lawful use. Moreover, since its establishment in 2021, the Defence AI Centre has played a key role in driving TEVV for military AI systems.

For example, in November 2025, it launched the AI Model Arena tool to help the MoD redefine its current approach to evaluating and procuring AI.¹²⁸ Finally, the Strategic Defence Review published in July 2025 has provided additional momentum through prioritising greater use of autonomy and AI.¹²⁹ While there are no recommendations within the document that are directly focused on TEVV for military AI systems, the proposed creation of a new defence research and evaluation organisation will also entail sustaining relevant TEVV capabilities within that organisation.¹³⁰

United States

In January 2026, the United States Department of War (DoW) published a memorandum outlining its new military AI strategy.¹³¹ The first heading in the strategy, ‘Accelerating America’s Military AI Dominance’, is aligned with the intent of Executive Order 14179 issued by US President Donald Trump in January 2025 on ‘Removing Barriers to American Leadership in Artificial Intelligence’¹³² and with the AI Action Plan issued by the White House in July 2025.¹³³ The DoW’s military AI strategy outlines four lines of effort: (1) experimenting with frontier AI models, (2) eliminating bureaucratic barriers to AI integration, (3) focusing investment in AI to leverage existing asymmetric advantages, and (4) carrying out ‘pace-setting projects’ (PSPs) to establish a new execution standard and build out the necessary foundation of enablers to accelerate AI integration.¹³⁴ Seven initial PSPs have been identified across warfighting, intelligence and enterprise mission areas: (1) ‘Swarm Forge’, focusing on discovering, testing and scaling with and against AI-enabled capabilities, (2) ‘Agent Network’, to support development and experimentation with AI agents, (3) ‘Ender’s Foundry’, focusing on AI-enabled simulation capabilities, (4) ‘Open Arsenal’, to accelerate the translation of technical intelligence into capability development, (5) ‘Project Grant’, to transform deterrence, (6) ‘GenAI.mil’, to broaden AI experimentation with the latest models, and (7) ‘Enterprise Agents’, focusing on AI agent development and deployment for enterprise workflows.¹³⁵ In addition to the memorandum on the DoW’s military AI strategy, two other memoranda were issued: one on restructuring the department’s Advana data platform¹³⁶ and the other on transforming the defence innovation ecosystem.¹³⁷ The latter memorandum brings all of the various defence innovation organisations under the leadership and responsibility of the Chief Technology Officer, namely the Under Secretary of War for Research and Engineering. Previously these organisations—the Defence Innovation Unit, Strategic Capabilities Office, Defence Advanced Research Projects Agency, Chief Digital and Artificial Intelligence Office (CDAO), Test Resource Management Centre (TRMC), and Office of Strategic Capital—were positioned at different levels of leadership and were part of various steering and working groups.

The DoW's Acquisition Transformation Strategy, published in November 2025, made three important changes to the department's approach to TEVV. First, it calls for modernising test infrastructure under the TRMC and creating a minimum set of department-wide test data and quality standards.¹³⁸ Second, it aims to reduce central TEVV oversight conducted by the Director, Operational Test & Evaluation (DOT&E), returning responsibility to the service branches.¹³⁹ Third, it calls for modernising systems engineering processes and tools informing TEVV 'to enable agile development, technology insertion, improved technology and manufacturing risk management, and reduced need for testing, rework, and re-testing to certify a system'.¹⁴⁰ In a similar vein, the DoW's military AI strategy specifically mentions adopting a '[w]artime approach to blockers',¹⁴¹ including in relation to TEVV. It also clarifies the DoW's approach to responsible military AI, which no longer emphasises ethical principles but rather focuses on what is lawful.¹⁴² As for the set of TEVV frameworks developed by the CDAO for military AI systems in 2024 under the previous administration, their status is unclear. These frameworks cover TEVV for AI models,¹⁴³ human–systems integration,¹⁴⁴ systems integration¹⁴⁵ and operational TEVV.¹⁴⁶ They are focused on providing guidance and best practices for formulating a TEVV strategy, rather than binding policy requirements or a step-by-step guide.

The United States' defence TEVV enterprise will continue to evolve as the DoW's new military AI and acquisition transformation strategies are implemented, but several aspects remain uncertain. Streamlining reporting lines for all the defence innovation organisations under the DoW's Chief Technology Officer and minimising central oversight by the DOT&E could produce a more dynamic and synergised TEVV enterprise that draws on a larger pool of capabilities and resources. However, there is also a risk of confusion over roles and responsibilities, which would increase rather than reduce friction. Furthermore, while plans to modernise TEVV infrastructure, processes and tools are uncontroversial, it is unclear whether this will leverage existing work such as the Joint AI Test Infrastructure Capability (JATIC) program initiated by the CDAO in 2023 with funding up to 2029¹⁴⁷ and the CDAO's TEVV frameworks published in 2024. Moreover, the existing DoW manual on TEVV for military AI systems (5000.101¹⁴⁸), which was last updated in December 2024, has yet to incorporate the policy changes that have taken place under the Trump administration since 2025. Of greater concern, however, is the sidelining of ethical principles in TEVV supporting responsible and lawful use of military AI systems. The United States did not endorse the recent REAIM Pathways to Action document, which adds to this concern. There is also the concern that the overall desire to ensure speed in the DoW's military AI strategy could mean that TEVV is marginalised.

Challenges

Six interrelated challenges associated with leveraging TEVV to measure trust and control in military AI systems were identified earlier: (1) determining the appropriate unit of analysis for TEVV, (2) the need for an AI system to have a clearly defined purpose for TEVV to evaluate against, (3) the impracticality of identifying all possible scenarios for TEVV, (4) the unpredictability and inscrutability of AI systems, (5) the fact that private sector TEVV tools and benchmarks may not neatly apply in the military context, and (6) the need for some parts of the TEVV process to be carried out continuously for AI systems that have models which are updated through data gathered during their use.¹⁴⁹ In addition to these challenges, the military domain imposes additional considerations that also affect TEVV. First, the scarcity of high-quality and representative data in operational environments will affect training datasets, the creation of representative test environments, and the ability to minimise bias.¹⁵⁰ Second, the deployment of military systems across different operational environments may necessitate retraining and re-evaluating when it comes to AI.¹⁵¹ Factoring in consideration of the complexities associated with human–machine teams further muddles what is already a very challenging set of circumstances for TEVV. Nevertheless, all of these challenges and considerations are known and therefore subject to ongoing work in terms of adapting TEVV processes and practices. This section builds on this by discussing challenges that have received comparatively less attention in the literature but are still important in terms of their impact on TEVV's role in supporting regulatory guardrails.

Lack of Standardised TEVV Practices

There are currently no standardised TEVV practices for AI in the military domain. While the nine-stage lifecycle framework developed by the IEEE Standards Association (see Figure 1) provides a foundation to build up a set of standardised TEVV practices, the fact that TEVV is both a stage within the lifecycle framework and an ongoing activity taking place throughout all stages already adds a layer of complexity. Moreover, the lack of standardised TEVV practices partly stems from the nature of AI systems and how AI is used in the military domain—particularly in terms of the ‘black box dilemma’, as well as in determining the unit of analysis and the system's purpose for TEVV, which are affected by how military systems can be deployed across different operational environments. Furthermore, where military AI systems are operating as part of human–machine teams, the ability to develop TEVV to account for real-world performance is limited not just to the technical or technological dimension but also to the human dimension.¹⁵² All of this points to a need for flexibility and adaptation in TEVV for military AI systems, which is precisely the dilemma in terms of the lack of standardised practices. Resolving this tension will be important to advance TEVV for military AI systems in such a way that it can then support responsible and lawful use of these systems while accounting for the practicalities of dealing with AI.

Across the case studies, it is clear that this is an ongoing challenge. In Australia's case, for example, if the service branches appear to be driving the overall approach to TEVV for military AI systems rather than Defence from a central standpoint, then we can already expect variations in practices based on the missions of each service branch. Similarly, for the United Kingdom, even though its MoD can leverage a common standard through JSP 936 and the Defence AI Centre's AI Assurance Framework, implementation at the service branch level is likely to necessitate some variations in TEVV practices. Turning to the United States, the minimisation of central oversight for TEVV by the DOT&E will mean that there will be a similar challenge in terms of managing the variation across service branches. The next section unpacks these points further in terms of factors shaping the implementation of TEVV, such as a military's technical capacity and resources, as well as its organisational, engineering and safety cultures. Additionally, the section on the private sector's involvement highlights another factor which further complicates the ability to develop standardised TEVV practices, namely the extent to which TEVV capabilities exist in-house within a military or depend on outsourcing to the private sector.

Variations in Implementing TEVV

Even if standardised TEVV practices can be identified, their implementation depends on a military's technical capacity and resources. These will vary considerably—among the case studies in this article, for example, there is a wide gulf between the United States and both Australia and the United Kingdom in terms of technical capacity and resources. The size of the US military's TEVV workforce was estimated at 10,000 in 2025, with 80 per cent being civilian personnel.¹⁵³ For Australia and the United Kingdom, a comparable estimate is much more challenging to make, as TEVV capabilities are distributed across civil servants, military personnel, and private sector contractors. In Australia's case, however, there was an estimated shortfall of 400 TEVV practitioners identified in 2024, which is expected to grow to 1,000 by 2030.¹⁵⁴ Estimating a military's investments in TEVV is also not a straightforward exercise, and a distinction must be made between what is allocated for TEVV within a specific capability project versus what is allocated at an organisational level. For example, in May 2025, the United Kingdom's MoD announced a £1.54 billion extension for its long-term partnering agreement with the British company QinetiQ to continue operating TEVV infrastructure and provide TEVV services.¹⁵⁵ In addition, the Defence and Security Accelerator announced a competition in November 2025 offering up to £1 million in funding for innovative TEVV capabilities as part of the MoD's Test and Evaluation Transformation Programme.¹⁵⁶ While significant, these developments do not represent the full extent of the MoD's investment in TEVV.

Moreover, how a military leverages its available technical capacity and resources will be shaped by its organisational, engineering, and safety cultures. For example, since the Trump administration took office in 2025, there has been a concerted attempt to

change the culture of the US military and defence bureaucracy, beginning with Executive Order 14347, issued in September 2025, which rebranded the Department of Defense as the Department of War.¹⁵⁷ In a speech to general and flag officers in the same month, Secretary of War Pete Hegseth spoke about restoring the ‘warrior ethos’ of the US military.¹⁵⁸ These changes have filtered down to shape the approach to TEVV for military AI systems. The DoW’s military AI strategy specifically states:

Diversity, Equity, and Inclusion and social ideology have no place in the DoW, so [it] must not employ AI models which incorporate ideological ‘tuning’ that interferes with their ability to provide objectively truthful responses to user prompts.¹⁵⁹

Moreover, the concern highlighted earlier regarding a desire to ensure speed in the DoW’s military AI strategy potentially leading to TEVV being marginalised is prompted by developments such as the rollout of GenAI.mil in December 2025, which appeared to be rushed, raising the question of whether there was an adequate level of assessment conducted.¹⁶⁰

Involvement of the Private Sector

Innovation in AI in the military domain is primarily being driven by the private sector within a handful of states. On the surface, this:

suggests fostering unrestrained private competence. On the other hand, there are strong incentives for state control, since military AI is a vital sector for national security, and many applications involve high security risks and significant political costs if left uncontrolled.¹⁶¹

There is a trade-off that states and militaries have to manage when it comes to ‘unconstrained expertise’¹⁶² that can support innovation but runs the risk of undesired outcomes, including the diffusion of innovative technologies to rival states. In the context of TEVV, because it is an important ongoing activity throughout the lifecycle of military AI systems (see Figure 1), navigating the trade-off between competence and control poses several challenges. First, it can significantly affect the timeline for capability development and acquisition. In a context where militaries are seeking to reform acquisition processes to keep up with the demands of AI, TEVV could be seen as a stumbling block. Second, the ability to develop in-house or sovereign TEVV capabilities depends on existing technical capacity and available resources. Even if states have the means to build sovereign TEVV capabilities, this is not something that can be achieved easily or within a limited amount of time. Finally, the lack of standardised TEVV practices limits the prospect of conducting TEVV independently, which reduces accountability.

The challenges arising from the competence–control trade-off are particularly evident in the case of the United States. For example, in March 2026, the DoW designated Anthropic as a ‘supply chain risk’, effectively restricting the use of its products and services by the US military and defence bureaucracy.¹⁶³ This unprecedented designation—typically reserved to guard against foreign companies from states deemed as adversaries—occurred due to a disagreement over Anthropic’s request to maintain safeguards over the use of its AI models for mass domestic surveillance and fully autonomous weapons.¹⁶⁴ Secretary of War Pete Hegseth attacked the company in a social media post, accusing it of using the ‘sanctimonious rhetoric of “effective altruism” ... [to] strong-arm the United States military into submission’.¹⁶⁵ Nonetheless, the DoW’s resort to such a measure is consistent with its military AI strategy, in which it states that the department must ‘utilize models free from usage policy constraints that may limit lawful military applications’.¹⁶⁶ The link to TEVV here is that the strategy directed the CDAO to ‘establish benchmarks for model objectivity as a primary procurement criterion within 90 days’. As of 10 April 2026, Anthropic is in the midst of legally challenging the ‘supply chain risk’ designation on the basis that it was denied due process, but US courts typically do not challenge executive branch decisions related to national security.¹⁶⁷

Conclusion

This article has aimed to explore the role of TEVV in supporting emerging regulatory guardrails for the responsible and lawful use of military AI systems. It began with an examination of the broader context for military AI, where facts on the ground are evolving much more quickly and at a scale that multilateral dialogue cannot keep up with. This was followed by an overview of TEVV across both the civilian and military domains, which provided a starting point to understand the complexities involved for the latter. These include consideration of practices for all military systems, as well as those for specific types of systems, and tailored practices at the service level, and for AI systems more generally, from a technical standpoint. Building on this, three specific areas were identified where TEVV supports regulatory guardrails: (1) providing the means to tangibly measure desired outcomes of an AI system, (2) enhancing human agency across the lifecycle of AI systems, and (3) enabling the conduct of legal reviews. However, this is not exhaustive—TEVV can also play a role in other ways, such as integrating expertise and assessments that were previously siloed, which can lead to more comprehensive and effective consideration of responsible and lawful use of military AI systems. Another example is the potential for TEVV to encourage cooperation and technical exchanges where such activities may have been constrained.

The three case studies examined—Australia, the United Kingdom, and the United States—demonstrate the range of possible ways to approach TEVV. They also highlight the tension between commitments made at platforms such as the REAIM summits and a state's pursuit of its interests. The case studies clearly demonstrate that, contrary to what its technical nature might suggest, TEVV is neither objective nor neutral—militaries leverage it in service of larger strategic objectives, and it is as much as a political endeavour as it is a technical one. The case studies also explore differences at the service branch level, and underscore the need for more detailed examination in future research of how differing operational, organisational and strategic contexts can have an impact on TEVV. While this article did not examine how TEVV shapes regulatory guardrails when it is conducted in the context of developing capabilities jointly as in a partnership like AUKUS, this is an additional area worth considering for future research. A further area worth exploring relates to developing strategies and approaches to address the three sets of challenges identified for TEVV's role in supporting regulatory guardrails: (1) the lack of standardised TEVV practices for military AI systems, (2) variations in implementing TEVV, and (3) the private sector's involvement. While it may be that some of the challenges cannot be addressed entirely, further discussion regarding what can be achieved will strengthen the ability of TEVV to facilitate responsible and lawful use of military AI systems.

About the Author

Manoj Harjani is a Research Fellow at the S Rajaratnam School of International Studies (RSIS), where he is also Coordinator of the Military Transformations Programme within the Institute of Defence and Strategic Studies. Prior to joining RSIS, Manoj served in roles focusing on analysing long-term trends and building capabilities in futures thinking and scenario planning with the Singapore Public Service. He holds a BA in Political Science and MA in South Asian Studies from the National University of Singapore.

ENDNOTES

- 1 Applications of AI in the military domain range across both operational activities (e.g. command, control and communications) and non-operational activities (e.g. acquisition and procurement). See Global Commission on Responsible AI in the Military Domain, *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (The Hague: Hague Centre for Strategic Studies, 2025), pp. 11–19.
- 2 See ‘TFL-1 Autonomy Modules: Terminal Guidance and Cruise’, *The Fourth Law* (website), at: <https://thefourthlaw.ai/#section4>.
- 3 Yuval Abraham, ‘“Lavender”: The AI Machine Directing Israel’s Bombing Spree in Gaza’, *+972 Magazine*, 3 April 2024.
- 4 Emelie Andersin, ‘The Use of the “Lavender” in Gaza and the Law of Targeting: AI-Decision Support Systems and Facial Recognition Technology’, *Journal of International Humanitarian Legal Studies* 16, no. 2 (2025): 336–370.
- 5 Tess McClure and Deepa Parent, ‘Minab School Bombing: How the Worst Mass Casualty Event of the Iran War Unfolded—A Visual Guide’, *The Guardian*, 3 March 2026.
- 6 Alex Woodward, ‘Old Intelligence and AI? Behind the Deadly Attack on an Iranian Girls’ School That Left 175 Dead’, *The Independent*, 12 March 2026.
- 7 Kevin T Baker, ‘AI Got the Blame for the Iran School Bombing. The Truth is Far More Worrying’, *The Guardian*, 26 March 2026.
- 8 For the list of endorsing states, see ‘REAIM: Responsible AI in the Military Domain Summit’, REAIM (website), at: www.exteriores.gob.es/en/REAIM2026/Paginas/default.aspx. The number of states endorsing the Pathways to Action document grew from 35 at the summit itself to 44 as of 31 March 2026. The text of the outcome document is REAIM Pathways to Action (A Coruña: Ministry of Foreign Affairs, the European Union and Cooperation, 2026), at: www.exteriores.gob.es/en/REAIM2026/Documents/REAIM%202026%20Pathways%20to%20Action.pdf.
- 9 Manoj Harjani and Mei Ching Liu, ‘The Uncertain Future of the REAIM Summit’, IDSS Papers, no. 029–26 (2026).
- 10 Ibid.
- 11 See *REAIM Call to Action* (The Hague: Government of the Netherlands, 2023).
- 12 See *REAIM Blueprint for Action* (Seoul: Ministry of Foreign Affairs, 2024).
- 13 See Bureau of Arms Control and Nonproliferation, *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* (US Department of State, 2023).
- 14 Ibid.
- 15 Valerie Hafez, Rania Wazir and Fariba Karimi, *AI as Complex Sociotechnical Systems: Problems, Approaches and Reflections* (Vienna: Women in AI Austria, 2023), p. 2.
- 16 Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI* (Geneva: United Nations Institute for Disarmament Research, 2020), p. 1.
- 17 Tim McFarland, ‘Reconciling Trust and Control in the Military Use of Artificial Intelligence’, *International Journal of Law and Information Technology* 30, no. 4 (2022): 472–483.
- 18 *REAIM Blueprint for Action*, p. 2.
- 19 Ibid., p. 2.
- 20 *REAIM Pathways to Action*, p. 3.
- 21 Ibid, pp. 4–5.
- 22 Bureau of Arms Control and Nonproliferation, *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*.
- 23 In the civilian domain, there is a rich body of standards associated with TEVV—for example, ISO/IEC/IEEE 24765:2017, ISO/IEC/IEEE 12207:2026, IEEE 829-2008, and IEEE 1012-2024.
- 24 Michael Borowski and Priscilla Glasgow, *When Systems Are Simulations—T&E, VV&A, or Both?* (McLean VA: MITRE Corporation, 1998), p. 1.

- 25 Feras A Batarseh , Laura Freeman and Chih-Hao Huang, 'A Survey on Artificial Intelligence Assurance', *Journal of Big Data* 8, no. 60 (2021): 2.
- 26 John Burden, Marko Tešić, Lorenzo Pacchiardi and José Hernández-Orallo, 'Paradigms of AI Evaluation: Mapping Goals, Methodologies and Culture', *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, Montreal, Canada, August 2025* (International Joint Conferences on Artificial Intelligence, 2025), p. 10382.
- 27 Ibid., p. 10384.
- 28 Ibid., p. 10386.
- 29 Ibid., p. 10386.
- 30 Bálint Gyevnár and Atoosa Kasirzadeh, 'AI Safety for Everyone', *Nature Machine Intelligence* 7 (2025): 531–542.
- 31 See 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023', GOV. UK, at: www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.
- 32 See 'Seoul Declaration for Safe, Innovative and Inclusive AI by Participants Attending the Leaders' Session of the AI Seoul Summit, 21st May 2024', *Ministry of Foreign Affairs, Republic of Korea* (website), at: www.mofa.go.kr/eng/brd/m_5674/view.do?page=1&seq=321007.
- 33 See 'Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet', *Permanent Mission of France to the United Nations in New York* (website), at: <https://onu.delegfrance.org/state-ment-on-inclusive-and-sustainable-artificial-intelligence-for-people-and>.
- 34 See 'AI Impact Summit Declaration, New Delhi', media release, 21 February 2026, *Ministry of External Affairs* (website), at: www.mea.gov.in/bilateral-documents?dtl/40809.
- 35 There is no universal definition for a military AI system. The term is used here in its broadest sense, encompassing standalone AI applications in the military domain as well as AI-enabled functionality within other types of military systems.
- 36 This includes practices from AI in other safety-critical areas, such as air traffic control, outer space, nuclear systems and medicine.
- 37 The relevant set of practices depends on the way in which a system is classified and defined. Given the challenges associated with defining military AI systems, this is also acknowledged as a challenge affecting TEVV.
- 38 Michael C Horowitz, 'AI and the Diffusion of Global Power', in Allison Leonard, Jennifer Goyder and Lynn Schellenberg (eds), *Modern Conflict and Artificial Intelligence* (Waterloo: Centre for International Governance Innovation, 2020), pp. 36–37.
- 39 RS Panwar, Li Qiang and John NT Shanahan (eds), *Military Artificial Intelligence Test and Evaluation Model Practices* (Geneva: INHR, 2024), pp. 4–5.
- 40 Heather M Wojton, Daniel J Porter and John W Dennis, *Test & Evaluation of AI-Enabled and Autonomous Systems: A Literature Review* (Alexandria VA: Institute for Defence Analyses, 2020), pp. 27–28.
- 41 Ibid., p. 19.
- 42 Manoj Harjani and Shantanu Sharma, *Adversarial Attacks: An Existential Threat to AI*, IDSS Paper No. 078 (RSIS, 2023).
- 43 McFarland, 'Reconciling Trust'.
- 44 Jean-Marc Rickli, Federico Mantellassi and Quentin Ladetto, *What, Why and When? A Review of the Key Issues in the Development of Military Human-Machine Teams* (Geneva: Geneva Centre for Security Policy, 2024), pp. 10–16.
- 45 McFarland, 'Reconciling Trust', p. 474.
- 46 Ibid., p. 477.
- 47 Heather M Roff and David Danks, "'Trust but Verify": The Difficulty of Trusting Autonomous Weapons Systems', *Journal of Military Ethics* 17, no. 1 (2018): 6.
- 48 Ibid., pp. 6–7.

- 49 Jovana Davidovic and Mitt Regan, 'Jus Ante Bellum and AI-Enabled Weapons', in Maria Power and Maggi Savin-Baden (eds), *Just War Theory and Artificial Intelligence: Challenges and Consequences* (Boca Raton FL: CRC Press, 2026), pp. 4–6.
- 50 Ibid., p. 4.
- 51 Ibid., p. 4.
- 52 Ibid., p. 5.
- 53 Ibid., p. 5.
- 54 Ibid., p. 5.
- 55 Ibid., pp. 5–6.
- 56 Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt and Patrick Hall, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* (Gaithersburg, MD: National Institute of Standards and Technology, 2022), p. 10.
- 57 *REAIM Call to Action*, p. 2.
- 58 Ibid., p. 2.
- 59 *REAIM Blueprint for Action*, p. 2.
- 60 Ibid., p. 4.
- 61 Bureau of Arms Control and Nonproliferation, *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*.
- 62 Ibid.
- 63 Ingvild Bode, Anna Nadibaidze, Tom Watts and Qiaochi Zhang, 'Ensuring the Exercise of Human Agency in AI-Based Military Systems: Concerns Across the Lifecycle', *Ethics and Information Technology* 27, no. 50 (2025): 2.
- 64 The AutoPractices Project, *Strengthening Human Agency in the Military Domain: Best Practices Toolkit for Policymakers, Developers, and Users of AI Systems* (Odense: Centre for War Studies, University of Southern Denmark, 2026), pp. 17–18.
- 65 *REAIM Blueprint for Action*, p. 2.
- 66 Ibid., p. 5.
- 67 Bureau of Arms Control and Nonproliferation, *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy*.
- 68 Klaudia Klonowska, *Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare*, Asser Research Paper 2021-02 (Asser Institute, 2021), p. 2.
- 69 Klaudia Klonowska and Jonathan Kwik, 'Artificial Intelligence in Contemporary Conflicts and the Future of Military Law', in Paul AL Ducheine, Terry D Gill, Peter BMJ Pijpers and Marten C Zwanenburg (eds), *A Research Agenda for Military Law* (Cheltenham: Edward Elgar Publishing, 2026), p. 105.
- 70 Tobias Vestner and Altea Rossi, 'Legal Reviews of War Algorithms', *International Law Studies* 97 (2021): 547.
- 71 Global Commission on Responsible AI in the Military Domain, *Responsible by Design*.
- 72 Ibid., p. 42.
- 73 Jonathan Kwik, *Iterative Assessment for Military Artificial Intelligence (AI) Systems*, ASSER Research Paper 2025-03 (Asser Institute, 2025), p. 14.
- 74 See 'The International Committee of the Red Cross as Guardian of International Humanitarian Law', *International Committee of the Red Cross* (website), at: www.icrc.org/en/article/guardian-international-humanitarian-law.
- 75 *Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach* (Geneva: International Committee of the Red Cross, 2019), p. 7.
- 76 Tom Corben and Sophie Mayo, *Federation Is Deterrence: The US Defence Industrial and Technology Integration Agenda in the Indo-Pacific* (Sydney: The United States Studies Centre, 2025).
- 77 See Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence* (Canberra: Commonwealth of Australia, 2025).
- 78 Ibid., p. 5.

- 79 Ibid., p. 6.
- 80 Ibid., p. 7.
- 81 Ibid., p. 8.
- 82 See 'Policy for the Responsible Use of AI in Government: Version 2.0', *digital.gov.au*, at: www.digital.gov.au/ai/ai-in-government-policy. The policy is, as of April 2026, in its second version since December 2025.
- 83 See 'Australia's AI Ethics Principles', *Department of Industry, Science and Resources* (website), at: www.industry.gov.au/publications/australias-ai-ethics-principles. Eight ethics principles were identified in this framework: (1) human, societal, and environmental wellbeing, (2) human-centred values, (3) fairness, (4) privacy protection and security, (5) reliability and safety, (6) transparency and explainability, (7) contestability and (8) accountability.
- 84 Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence*, p. 9.
- 85 Ibid., p. 9.
- 86 See Department of Defence, *Defence Test & Evaluation (T&E) Strategy* (Canberra: Commonwealth of Australia, 2021).
- 87 Ibid., p. 3.
- 88 Department of Defence, *SDIP 7—Test and Evaluation, Certification and Systems Assurance* (Canberra: Commonwealth of Australia, 2024), p. 4.
- 89 See Australian Defence Force, *Concept for Robotic and Autonomous Systems* (Canberra: Commonwealth of Australia, 2020).
- 90 Ibid., p. 31.
- 91 Ibid., p. 35.
- 92 Ibid., p. 59.
- 93 Australian Public Service Commission, *State of the Service Report 2024–25* (Canberra: Commonwealth of Australia, 2025), p. 112.
- 94 See Army, *Robotic & Autonomous Systems Strategy 2.0* (Canberra: Commonwealth of Australia, 2022).
- 95 Ibid., p. 32.
- 96 Ibid., p. 33.
- 97 Ibid., p. 34.
- 98 Ibid., p. 36.
- 99 See Navy, *RAS-AI Strategy 2040* (Canberra: Commonwealth of Australia, 2020).
- 100 Ibid., p. 19.
- 101 Ibid., p. 24.
- 102 Ibid., p. 24.
- 103 See Royal Australian Air Force, *Plan Jericho* (2021), at: <https://airpower.airforce.gov.au/sites/default/files/2021-03/AF14-Plan-Jericho.pdf>.
- 104 See 'Air Warfare Centre', *Royal Australian Air Force* (website), at: www.airforce.gov.au/about-us/hq-air-command/air-warfare-centre.
- 105 See 'M113 AS4 Optionally Crewed Combat Vehicle', *BAE Systems* (website), at: www.baesystems.com/en/product/m113-occv.
- 106 See 'Bluebottle: Satellites of the Sea', *OCIUS* (website), at: <https://ocius.com.au/usv/#overview>.
- 107 See 'MQ-28 Ghost Bat', *Boeing* (website), at: www.boeing.com/defense/autonomous-and-unmanned-systems/mq-28-ghost-bat.
- 108 'AI Takes Flight on Talisman Sabre', *Department of Defence* (website), 11 August 2025, at: www.defence.gov.au/news-events/news/2025-08-11/ai-takes-flight-talisman-sabre.
- 109 'AUKUS Pillar II Under Pressure', *Strategic Comments* 31, no. 35 (2025): 1–4.
- 110 Ministry of Defence, *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence* (London: Ministry of Defence, 2022), p. 1.
- 111 Ibid., pp. 10–11.
- 112 See Ministry of Defence, *Defence Artificial Intelligence Strategy* (London: Ministry of Defence, 2022).

- 113 Ibid., p. 13.
- 114 See Ministry of Defence, *JSP 936 v1.1 Dependable Artificial Intelligence (AI) in Defence Part 1: Directive* (London: Ministry of Defence, 2024).
- 115 See Ministry of Defence, *JSP 375 Management of Health and Safety in Defence* (London: Ministry of Defence, 2013).
- 116 See Ministry of Defence, *JSP 376 Defence Acquisition Safety Policy* (London: Ministry of Defence, 2023).
- 117 See Ministry of Defence, *Defence Information, Knowledge, Digital and Data Policy Commitments* (London: Ministry of Defence, 2024).
- 118 See Ministry of Defence, *JSP 912 Human Factors Integration in Defence Systems* (London: Ministry of Defence, 2014).
- 119 See Ministry of Defence, *JSP 939 Defence Policy for Modelling and Simulation* (London: Ministry of Defence, 2018).
- 120 See Ministry of Defence, *Test and Evaluation (T&E): Future Advantage Through Evaluation (FATE)* (London: Ministry of Defence, 2026).
- 121 Ministry of Defence, *Defence Artificial Intelligence Strategy*, p. 27.
- 122 *JSP 936 V1.1 Dependable Artificial Intelligence (AI) in Defence Part 1*, p. 30.
- 123 See 'The Defence AI Centre (DAIC) AI Assurance Framework', *Ministry of Defence* (website), at: www.digital.mod.uk/policy-rules-standards-and-guidance/aiph.
- 124 See 'AI Practitioner's Handbook (AIPH)', *Ministry of Defence* (website), at: www.digital.mod.uk/policy-rules-standards-and-guidance/aiph.
- 125 See 'Resource Library', *Ministry of Defence* (website), at: www.digital.mod.uk/policy-rules-standards-and-guidance/aiph/resource-library.
- 126 See Ministry of Defence, *British Army's Approach to Artificial Intelligence* (London: Ministry of Defence, 2023).
- 127 House of Commons Defence Committee, *Developing AI Capacity and Expertise in UK Defence* (London: House of Commons, 2025), pp. 26–29.
- 128 Ministry of Defence, 'Launching the AI Model Arena', *GOV.UK*, at: www.gov.uk/government/news/launching-the-ai-model-arena.
- 129 Ministry of Defence, *Strategic Defence Review: Making Britain Safer: Secure at Home, Strong Abroad* (London: Ministry of Defence, 2025), p. 96.
- 130 Ibid., 56.
- 131 See Secretary of War, *Artificial Intelligence Strategy for the Department of War* (Washington DC: Department of War, 2026).
- 132 See President, *Removing Barriers to American Leadership in Artificial Intelligence*, Executive Order 14179 (Executive Office of the President, 2025).
- 133 See *Winning the Race: America's AI Action Plan* (Executive Office of the President of the United States, 2025).
- 134 Secretary of War, *Artificial Intelligence Strategy for the Department of War*.
- 135 Ibid., pp. 2–3.
- 136 See Deputy Secretary of War, *Transforming Advana to Accelerate Artificial Intelligence and Enhance Auditability* (Washington DC: Department of War, 2026).
- 137 See Secretary of War, *Transforming the Defense Innovation Ecosystem to Accelerate Warfighting Advantage* (Washington DC: Department of War, 2026).
- 138 U.S. Department of War, *Acquisition Transformation Strategy* (Washington DC: Department of War, 2025), pp. 21–22.
- 139 Ibid., p. 22.
- 140 Ibid., p. 29.
- 141 Secretary of War, *Artificial Intelligence Strategy for the Department of War*, p. 4.
- 142 Ibid., p. 5.

- 143 See Chief Digital & Artificial Intelligence Office (CDAO), *Test and Evaluation of Artificial Intelligence Models* (CDAO, 2024).
- 144 See Chief Digital & Artificial Intelligence Office (CDAO), *Human Systems Integration Test and Evaluation of Artificial Intelligence-Enabled Capabilities* (CDAO, 2024).
- 145 See Chief Digital & Artificial Intelligence Office (CDAO), *Systems Integration Test and Evaluation of Artificial Intelligence-Enabled Capabilities* (CDAO, 2024).
- 146 See Chief Digital & Artificial Intelligence Office (CDAO), *Operational Test and Evaluation of Artificial Intelligence-Enabled Capabilities* (CDAO, 2024).
- 147 See 'CDAO JATIC Documentation', CDAO (website), at: <https://cdao.pages.jatic.net/public>.
- 148 See Department of Defense, *DOD Manual 5000.101 Operational Test and Evaluation and Live Fire Test and Evaluation of Artificial Intelligence-Enabled and Autonomous Systems* (Department of Defense, 2024).
- 149 Davidovic and Regan, 'Jus Ante Bellum and AI-Enabled Weapons', pp. 4–6.
- 150 Panwar, Qiang and Shanahan (eds), *Military Artificial Intelligence Test and Evaluation Model Practices*, p. 6.
- 151 Ibid., p. 6.
- 152 David Helmer, Michael Boardman, S Katie Conroy, Adam Hepworth and Manoj Harjani, 'Human-Centred Test and Evaluation of Military AI', *arXiv* (2024), p. 6, at: <https://arxiv.org/abs/2412.01978>.
- 153 Office of Systems Engineering and Architecture, *Test and Evaluation Workforce Report* (Washington DC: Office of the Under Secretary of Defence for Research and Engineering, 2025), p. 9.
- 154 Department of Defence, *SDIP 7—Test and Evaluation, Certification and Systems Assurance*, p. 3.
- 155 See *T3E* (website), at: www.t3e.uk/en.
- 156 Defence and Security Accelerator, 'Delivering Future Advantage Through Testing and Evaluation: New £1 Million Themed Competition Launched', *GOV.UK*, at: www.gov.uk/government/news/delivering-future-advantage-through-testing-and-evaluation-new-1-million-themed-competition-launched.
- 157 See President, *Restoring the United States Department of War*, Executive Order 14347 (Executive Office of the President, 2025).
- 158 'Secretary of War Pete Hegseth Addresses General and Flag Officers at Quantico, Virginia', speech, Quantico VA, 30 September 2025, transcript at: www.war.gov/News/Transcripts/Transcript/article/4318689/secretary-of-war-pete-hegseth-addresses-general-and-flag-officers-at-quantico-v.
- 159 Secretary of War, *Artificial Intelligence Strategy for the Department of War*, p. 5.
- 160 Brandi Vincent and Drew F Lawrence, 'The Era of GenAI.mil Is Here. Users Have Mixed Reactions and Many Questions', *Defense Scoop*, at: <https://defensescoop.com/2025/12/18/genai-mil-users-have-mixed-reactions-and-many-questions>.
- 161 Andrea Johansen and Andreas Kruck, 'The Competence–Control Trade-Off in Military AI Innovation: Autonomous Weapons Systems and Shifting Modes of State Control over Private Experts', *European Journal of International Security* (2025): 3.
- 162 Ibid., p. 6.
- 163 Karen Freifeld and Deepa Seetharaman, 'Pentagon Designates Anthropic a Supply Chain Risk', *Reuters*, 6 March 2026.
- 164 'Statement from Dario Amodei on Our Discussions with the Department of War', *Anthropic* (website), 26 February 2026, at: www.anthropic.com/news/statement-department-of-war.
- 165 Pete Hegseth (@SecWar), 'This week, Anthropic delivered a master class in arrogance and betrayal as well as a textbook case of how not to do business with the United States Government or the Pentagon', X, 28 February 2026, at: <https://x.com/SecWar/status/2027507717469049070>.
- 166 Secretary of War, *Artificial Intelligence Strategy for the Department of War*, p. 5.
- 167 Jack Queen, 'Anthropic Has Strong Case Against Pentagon Blacklisting, Legal Experts Say', *Reuters*, 11 March 2026.

/ DIGITISING THE COMMANDER'S INTENT TOWARDS EFFECTIVE GOVERNANCE OF DECENTRALISED HUMAN–AI TEAMS

Adam Hepworth and Hussein Abbass

Introduction

Commander's intent (CI) is a directive framework under uncertainty that promotes action and initiative in military operations, functioning as a semantic compression mechanism across teams with shared situational and contextual understanding. It enables disciplined initiative through the delegation of authority, where the commander remains accountable for mission outcomes. Shared understanding, built through professional military education, common doctrine, individual and collective training, and trust developed through the command climate, has enabled CI to operate effectively in predominantly homogeneous human teams. The introduction of AI-enabled autonomous systems (AIAS) into military operations begins to fracture the underlying cognitive and physical homogeneity assumptions of the mission command paradigm.

In our prior work, we introduced meaningful human command (MHC1) as an evolution of meaningful human control (MHC) for military human–robot interaction (HRI). In this paper, we address an enabling requirement for MHC1 that our prior work did not specify: a technical and governance mechanism for disciplined initiative under uncertainty in human–machine teams, which we propose as the digital commander's intent (DCI). DCI marks a transformation from an action-oriented view of CI to a governance-oriented view in which CI itself becomes the instrument that establishes purpose, boundary conditions and risk appetite for human and machine subordinates. We articulate the problem space, identify three challenges for extending CI to mission command in human–machine teams, propose a four-skill framework for commanders and outline a pathway for future work.

Commander's Intent in Human Settings

The employment of AI-enabled autonomous systems (AIAS) in military operations is increasing globally, challenging existing governance and systems of control.¹ Existing international legal² and operational³ frameworks build on the assumption that a human commander retains decision custodianship throughout all phases of an operation. The introduction of AIAS challenges this paradigm, in particular the underlying ability of a commander to assess, direct and verify the actions of *commandeers*—*any delegatee*, be they human or machine.⁴ This paper addresses the implications of extending the concept of commander's intent (CI) across heterogeneous human-machine teams and the associated governance challenges of fulfilling command responsibilities as decision-making boundaries become less well defined. Mission commanders oversee the governance framework of the mission. Commanders must be conscious that their intent will interact with governance controls, especially in environments with automated workflows such as those involving AIAS entities. Because a commander's intent must operate through these governance controls, a competent AIAS may try to relax a specific control to satisfy the CI—that is, a capable AIAS pursuing the commander's intent may trade away a governance control to reach the goal. The digital commander's intent (DCI) prevents this: it makes explicit which controls are binding and which may flex, and by how much.

In our prior work, we established meaningful human command (MHC1) as an evolution of meaningful human control (MHC) for military human-robot interaction (HRI) settings.⁵ This paper addresses an essential mechanism that MHC1 requires but did not specify: a formal representation of CI for heterogeneous human-machine teams.

Commander's Intent: Definition and Role

The Australian Defence Force defines the commander's intent (CI) as a 'clear and concise expression of the operation's purpose, method and end state'.⁶ The United States Civil Air Patrol Unit Commanders Course guide describes a CI as 'the description and definition of what a successful mission will look like'.⁷

Expanding the aperture to the link between CI and mission command, UK doctrine asserts that mission command is 'founded on the clear expression of intent by commanders and the freedom of subordinates to act to achieve that intent', expressing that 'trust is a prerequisite of mission command' and going on to highlight that 'commanders will decide the extent of delegation based upon the specific context and judgement of their subordinates'.⁸ Canadian doctrine positions mission command as a key to the manoeuvrist approach, articulating that it is 'the philosophy of command that promotes unity of effort, the duty and authority to act, and initiative to subordinate

commanders'. Three key tenets are elucidated: the importance of understanding superior intent; clear responsibility of subordinates to achieve the intent; and decision-making in a timely manner.⁹

These diverse national perspectives jointly surface a common conceptual connection, the CI, being anchored in the mission command philosophy: the purpose is to succeed on all criteria that matter to the commander. It is a natural expression and expectation of effective decision-making to describe what success looks like, what is acceptable, what is unacceptable, and what the conditions and failure modes are—that is, the no-go conditions and the threshold the commander establishes to differentiate success from failure. CI is the minimum description of purpose and the goal that remains informative regardless of how uncertainty in the operational environment changes. It is the invariant *why* and the desired end state that remains when tasks, synchronisation and uncertainty invariably require adaptation to changing operational contexts.

The practice of command exists on a spectrum ranging from centralised to decentralised control models, with each extreme representing a model with unique characteristics and features.¹⁰ CI enables a disciplined form of delegation under a decentralised model, in which a commander delegates to at least one commandeer. These commandeers, in return, are expected to take initiative and act within the bounds of the rules of engagement (ROE), military values and legal boundaries. As uncertainty escalates, the complexity or scale of a mission increases and commanders must delegate more, transitioning from hands-on to hands-off execution. This form of delegation, enabled and exercised through mission command (MC), is imperative when operations face severe uncertainties in contested environments. It provides CI at all levels where command and control (C2) is practised, from a military section's second-in-command to strategic command through government to the most senior military leaders. CI enables both individuals and teams to exercise disciplined initiative; the orders become a framework to operate within, in contrast to a specific set of 'only do' actions.

Exercising Mission Command

Mission command (MC) represents the practical upper-bound expression of decentralised execution, where commanders delegate authority to commandeers who subsequently perform command functions within the delegated scope.¹¹ While commanders delegate authority and responsibility for execution in MC, a commander cannot delegate accountability.¹² The commander remains accountable for mission success, invariant to the degree of delegation, including ancillary impacts of mission performance, in addition to achievement of mission objectives. The ability to delegate execution requires significant skills from commanders. CI as a mechanism operationalises the concept of MC, offering a path for commanders to enable commandeers to act in situations of physical dislocation, communications degradation or both.

An underlying and often unstated assumption of MC is that the CI framework ‘works’ for human teams with its inherent semantic compression¹³ due to a common lexicon and training paradigm, shared concept space and inculcated culture of practice. Due to this shared space, there is an explicit part of CI and an implicit one,¹⁴ in which commanders use their shared understanding of the doctrines to infer that their plans and actions will align with the commander’s overall intent.¹⁵

The fundamental idea of MC holds when transitioning from homogeneous human teams to heterogeneous human–machine teams; however, implementations begin to fracture as the complexity of command increases. Such underlying assumptions embedded as elements of pre-specified intent may require further shared awareness and specification, with a commander needing to adapt the CI as the situation evolves. For example, as these assumptions begin to fracture in human teams operating with partners from diverse linguistic perspectives, different national backgrounds and divergent cultural understandings, CI as an operating modality can become brittle in practice. The introduction of robotic team members further increases team heterogeneity and amplifies this brittleness, both cognitive and physical, with differences becoming more apparent as assumptions of the underlying model fracture and fail.¹⁶ The shared understanding is decoupled among agents unless explicitly engineered. In such cases, a new construct is necessary to highlight and bound the effect of hidden assumptions in both the explicit and implicit parts of a CI, contributing to team assurance across heterogeneous teams and systems.

Commander’s Skills

Research on MC has identified factors of success, including necessary enabling conditions. Recent work suggests that MC is a cultural and institutional phenomenon for the Australian Army.¹⁷ While not the focus of this article, we must briefly highlight a few critical commander’s skills that will drive the foundations for our discussion on CI for AIAS: commanders’ capacity assessment, commander’s risk foresight, clarity of communication and commander’s comprehension.

1. **Commanders’ capacity assessment:** Commanders issuing direction need to have the ability to assess the capacity of commanders receiving direction. Commanders are accountable, and are therefore responsible for ensuring that the delegatee has the capacity to perform the delegated responsibility. Capacity here includes the delegatee possessing the skills and competencies to perform and lead the delegated tasks—that they have the cognitive and physical resources and are responsible moral agents bounded by law, obligations to the welfare of soldiers, and military values. For AI-enabled commanders, we posit that test, evaluation, verification and validation (TEVV) is analogous to the qualifications, competency and currency of homogeneous

human team members. It follows that the need for a responsible human agent increases, with the AIAS unable to bear a moral responsibility. The commander's ability to assess the capacity of the commandeers through this lens must necessarily be calibrated to the bounds of machine performance, both desirable and undesirable.

2. **Commander's risk foresight:** When commanders maintain close control of their commandeers in simple missions, they are directly exposed to the situations their commandeers face and can assist in managing unfolding risks. As commanders step back and delegate, distancing themselves from the situations their commandeers face, they need to be skilled in risk foresight for the situations their commandeers will face. These skills could be developed in simulated settings to accelerate commanders' training; these must include scenarios to improve the commander's competencies for assessing the psychological, physical and mission stressors that their commandeers face. Mission stressors may impact the capacity of the performer, increase uncertainty for commanders, degrade team culture and impact mission success. For AIAS, the concept of the commander's risk foresight must generalise to include known and plausible machine failure modes, calibrating behavioural envelopes beyond procedural controls. The key task for the commander in such a situation is the same as with human commandeers: to identify the boundary conditions under which the envelope no longer holds. The commander must understand changes in the situation over time, particularly the assumptions underpinning risk tolerances and when risks deviate beyond the established risk appetite, necessitating reassessment.
3. **Clarity of communication:** Communication has been a cornerstone in the discussion of command. Commanders are expected to have clarity about what they want commandeers to achieve; they need to communicate their CI clearly. This expectation is managed in a human-only setting through education and training. In the military, standards in training regimes can achieve scale through training the whole organisation on concrete linguistic structures, including lexicons, to be used and therefore understood by everyone. While the style of communication in the military has been successful in delivering military missions, it is well documented in the literature that implementation of MC varies between doctrine and practice.¹⁸ For example, there is a body of literature on the communication challenges that arise from civil–military cooperation.¹⁹ Similarly, communication styles in academia vary by discipline, posing challenges for multidisciplinary research. Moreover, the linguistic nuances of spoken language and freeform documents, while offering flexibility in communication, introduce complexities in interpretation and in the resulting mental models. A structured document with clear linguistic structure and rules could reduce some of these effects through education and transfer learning—for example, from other structured military documents such as task orders.

4. **Commander's comprehension:** In a human team context, a commander requires the necessary intuition to 'read' how commandeers are performing during execution of tasks. This allows the commander to evaluate how the execution is progressing—or, if it is diverging, why. Seminal work in the field discusses this in the three situational awareness (SA) levels model, featuring Level 1 (perception), Level 2 (comprehension) and Level 3 (projection), extending beyond awareness and understanding to anticipation of future states.²⁰ Within human teams, this anticipatory skill is enabled by shared cognitive and cultural contexts. For heterogeneous human–machine teams, a commander's skill degrades as heterogeneity increases, due to distinct failure modes and new signals to learn, necessitating the development of new competencies to interpret outputs. In our setting, we position 'comprehension' after anticipation and communication, distinct from established framing and language in the literature. This deliberate choice reflects the nested cycles of situation awareness, in which the first three skills enable the commander to operate effectively within the team and the fourth skill operates at a higher level of abstraction, synthesising team-level outputs into an integrated assessment of mission and system-of-systems state. Distinct from Level 2 SA, which precedes projection, our fourth skill encompasses all three SA levels applied at the mission scale and is predicated on functioning team-level SA.

Functional Elements of CI

The practical implementation of CI is as an instrument, often delivered through a document, for decentralised work execution under uncertainty. It provides a stable core of direction, including purpose ('why'), tasks and effects ('what to do' and 'what not to do') and the end state ('what to accomplish').²¹ The upshot here is a framework that provides operational resilience as the context and plans change, with intent remaining static until revised by the commander. It is important to understand the linguistic and functional decomposition described above, and equally important to appreciate the interdependencies and intricacies of CIs. A CI, therefore, is not a wish list that commanders aspire to achieve. A CI describes success on one level such that, when all CIs on this level are successful, they describe elements of success for higher levels of command. The ADF Command philosophical doctrine states:

[T]he commander's intent provides the unifying idea that allows decentralised decision-making and execution within an overarching framework. Understanding the commander's intent two levels up enhances unity of effort by ensuring those on the ground have an understanding of the bigger picture as it applies to them.²²

This nesting requires alignments across all levels of a hierarchy, affording an understanding of how they are situated within the broader operation. Different literature presents complementary structures on the contents of a CI. The variations are predominantly due to differences of scale, organisational structure, scope and mission types. The constant information needed in any CI answers the following three questions:

1. What is the current situation (why is the mission needed)? This could be a strategic, tactical or operational situation as an element of the broader context.
2. What is the mission? This could include mission scope, resources, and rules of engagement, as well as additional essential elements of information.
3. What does success look like? This could vary between a quantitative description of success, such as delivering all cargo to their destinations (homogeneous weighting), or a description such as delivering all cargo that is of high military value (heterogeneous weighting). A description of success could also set constraints on operations such as safety, ethical boundaries and legal boundaries. Success could also be defined qualitatively, such as 'protect the force' or 'delay the enemy as much as possible', where neither description has a concrete objective form or a quantitative threshold to define what 'protect' and 'as much as possible' mean.

Command Workflow

A CI is only a single document in a mission workflow. It exists in a broader context of orders, guidance, constraints, rules and command direction. A CI could be the starting point in a workflow that then informs one or more task orders, which in turn gets decomposed into other sub-task orders for different teams. Similarly, the CI needs to be tested with the CI of the commander two levels up. A CI delivers information that propagates through the workflow; thus, understanding the overall workflow is vital, especially in the context of MC. It is inadequate to describe only the CI format or document structure without understanding how the information embedded in the CI will propagate into the rest of the workflow, the nature and type of the audience who need to put it into action, and how the information needs of the workflow influence and shape the contents of the CI.

The complexity of the workflow depends on the scope of the CI. For example, when a strategic leader issues a CI, the scope may be an entire nation or organisation, whereas when a tactical commander issues a CI in a tactical mission, the scope is limited to the boundaries of the mission and the personnel with a need to know. Integrating workflows across a large operation or an organisation must necessarily include the scope of the tactical commander's mission. It is important in this form of integration not only that the classic information for workflow—including task orders—is integrated but also that intents are integrated and linked to the overall effect space of the organisation's jurisdiction.²³

A commander exists for a purpose. In a more holistic context, a commander delivers the effects expected to meet strategic and national direction. This strategic statement then propagates down through the command hierarchy and is decomposed into intents that may linguistically seem distant from the strategic intent but functionally are the vehicles necessary to deliver it. For example, a tactical commander may not describe the strategic intent in a CI statement; instead, they may focus their CI on operational or higher-level tactical intents. In a well-designed organisation, the intent of a commander takes the form of an executive order to deliver the intent of a possibly more senior commander. The nesting of a CI's 'two-up' ensures semantic connectedness and mission synchronisation at all levels of command to form the basis for governance in the C2 system as the connective tissue of the organisation.

The inclusion of AIAS in this integrated workflow suggests two requirements. The workflow must be digitised and automated to enable seamless coordination and tasking of AIAS. The workflow must also rely on a common language for both humans and AIAS to communicate information or have the capacity to translate between languages with sufficient fidelity to preserve intent. We posit that translation is prone to error due to linguistic inequivalence and ambiguities, multiple interpretations when one language is transformed to another, and the underlying fact that AIAS do not receive the classically grounded information that machines with low levels of autonomy require.²⁴ Provided with explicit bounds, MC has the potential to expose heterogeneous human–machine teams to brittleness during the exercise of 'initiative'.²⁵ If the CI remains a governing construct for human–machine mission command, the explicit form of translation into machine action will need to consider how to transform intent into an objective in a context.

Transitioning from Humans to Machines—Requirements and Challenges

Towards a Digital Commander's Intent

We now transition our discussion from CI as an instrument of command aimed at delivering a directive expression of purpose, method and end state, to CI as an instrument of governance. The CI has classically been communicated from the human commander to the human commandeers. Machines, automation and digital decision-support systems have been tools in this human-centric decision-making cycle, in part due to the low cognitive readiness for computer-based systems to compute, make decisions and have a degree of freedom in executing actions when faced with problems under uncertainty. As the processing capacity of AIAS to explore large option spaces in support of CI increases, its cognitive readiness—and, consequently, its autonomy level—increases too.

The increase in autonomy readiness can lead to an increase in the complexity of tasks assigned to these systems and the degrees of freedom to generate, evaluate, execute, monitor and assess diverse courses of action (COAs).

Digitisation in this sense is the deliberate act of drawing out implicit assumptions and treating them as variables to be considered, particularly due to growing cultural and physical divergences between human and non-human cognitive agents (i.e., disparate cultural and contextual perspectives). We suggest that the establishment of a DCI must extend practices undertaken between human–human teams of distinct cultural backgrounds to generate a new shared understanding. As a working definition in this paper, we suggest that DCI is given as a formal representation of CI designed for bidirectional communication and shared understanding across heterogeneous human–machine teams. MC with human–human teams could be broadly thought about as treating understanding as observable through behaviours, composed of actions and information. More abstractly, actions required by tasks and ends shaping the effect space can become the evaluation framework to decompose how a machine may be planning and conducting a task.

Most machines in their current autonomy-readiness level are incapable of handling even a constrained form of a CI. If CI is seen as the invariant why, with a task order being seen as the concrete what and descriptive how, it is important to map out some of the fundamental requirements for a machine to handle a (constrained form of) CI. In this paper, we will assume advanced AIAS systems capable of performing missions autonomously—that is, once delegated a mission from the commander, the AIAS can execute it autonomously without further communication with the commander, except when necessary and possible. We will refer to these as DCI-ready systems. Through this lens, the DCI is a general form of CI that relaxes homogeneous cognitive and physical assumptions.

Contemporary CI can be communicated as a static artefact, from higher-level commanders to lower-level commanders that possess a shared lexical, cultural and semantic frame. This enables the interpretation of ambiguity through a common contextual lens, often leading to an end state that is within the realm of the commander's desired (or expected) state. However, where this may begin to break down is with the introduction of AIAS, as a result, in part, of the divergence between lexical and semantic frames across human and non-human cognitive agents.

Challenge 1: Minimum Requirements for Machines to Be Issued with a CI

A DCI-ready system is assumed to have explicit decision loops formulated around a version of the observe–orient–decide–act (OODA) framework.²⁶ However, OODA is not merely an architecture for the machine but an architecture for what humans need to understand about the machine in a command cycle. Humans may ask questions about what the machine observed; how the machine integrated and adjusted these observations within background contextual information; how COAs were generated; what criteria were used to decide the best COA; and when, where and how the machine used the chosen COA to actuate on itself and the environment. Each phase of OODA should have the capacity to be queried as a result of traceability, auditing, accountability and situation-awareness expectations, and the ability to answer using an appropriate reasoning approach that does not compromise need to know and other safety and security requirements.

Underpinning the above requirement for an OODA-interoperable reasoning box, both humans and machines need to exchange messages efficiently. Transparency requirements of interpretability and explainability are necessary; however, it is important to emphasise that the core requirement is understanding. Both humans and machines need to understand the information they receive. To effectively manage the high-tempo nature of military operations, this understanding needs to occur using minimum information—that is, messages need to be semantically compressed, shared for integration with previous messages to avoid redundancies, and considered within the bandwidth of communication channels and cognitive actors. Both machines and humans have different, but ultimately limited, bandwidth. For a human, this is known as cognitive capacity; a machine is the same, with limited memory and processing capacities. While transparency is necessary, it is not sufficient for understanding.²⁷

Thus, the DCI requires explicit representation designed for bidirectional human–machine communication, in contrast to approaches dominant today that adapt concepts for intent sharing from human natural language. This thinking and approach build on existing literature focusing on structured communication languages, in which the objective is not linguistic realism but rather the reduction of information ambiguity to reduce coordination and reliability uncertainty.²⁸ This draws out a new nuance in the trade space: while constraining language increases traceability and assurance, it may also risk oversimplifying the underlying values inherent in CI. The DCI must preserve essential information (constraints, risk appetites and tolerances, and bounds) while explicitly articulating what cannot be formalised (situational atmospherics and contextual nuance).

The last set of requirements for a machine to be ready to process a (constrained form of) CI is related to those stemming from the disciplined initiative expected in a commandeer executing MC; these include policies, processes, procedures, doctrines, and legal and ethical requirements. The transition between ‘hands-on’ and ‘hands-off’ execution necessitates the commander having the ability and capacity to assess a commandeer’s capacity. With AIAS, this must be both analytical and technical, assessing not only the ‘capacity’ of battery life or system range but—arguably more importantly—its cognitive resource capacity and biases. This could include such aspects as (un)certainty, data relevance (training to mission divergence), functions and the impact of different parameterisations, and ethical alignment. Recall that the commander retains accountability.²⁹

It is therefore necessary to calibrate the verified performance envelope of the machine, with CI constraints modulated relative to ability. Let us take two boundary cases as an example. In the low-capacity case, to meaningfully address the scenario (for instance) where a novel situation is beyond the training distribution of the machine, a detailed (prescriptive command) approach may be taken with explicit and controlled bounds. In the high-capacity case, a guided approach (MC) may be assumed with loose constraints and a dynamic state space for allowable actions, changing dynamically without reauthorisation in response to the necessity of the scenario.³⁰

Challenge 2: CI as an Instrument of Governance in Military HRI

As a practical element of human–machine teaming, we consider the CI as an instrument of governance, establishing purpose, boundary conditions, and the risk appetite and tolerances that a commandeer agent—be they human or machine—may exhibit as the context changes and actions deviate from the agreed plan. In contrast to more abstract governance artefacts, CI in this setting underwrites key principles of mission command, such as disciplined initiative, enhancing accountability logic.³¹

Determining a context-appropriate level of CI for AIAS requires determining whether to translate high-level human intent from a human to a machine or extend specific machine orders to humans. Generalisation may be a promising path of exploration, based on the underlying assumption that to achieve MHC1,³² we must extend beyond constrained procedural control paradigms, enabling machine exploration of new subspaces and the generation of novel COA. This assumption parallels how CI works for human teams; however, it requires further validation. Such an approach affords a ‘task order’ with flexible boundaries, as with human agents and teams, in contrast to a state vector of static information, as a decomposition of intent from human expressiveness to machine specificity.

This approach necessitates further consideration of accountability mechanisms, transforming from an action-oriented view (CI) towards a governance-oriented view (DCI). This opens an opportunity to consider new ways of addressing existing legal and governance challenges around international humanitarian law,³³ ensuring consistent command accountability³⁴ and decision custodianship. Where contemporary implementations of CI rely on several implicit assumptions (such as shared understanding) to maintain accountability, the governance-oriented view exposes the delegatory bounds of parameterised authority, including risk tolerances, constraints and ROE. These are necessary conditions when the commander is an AIAS, operating beyond direct human supervision and oversight. The detailed design for a DCI as a governance architecture must consider its interaction with legal and ethical obligations, as well as command responsibility.

Challenge 3: Commander's Skills for Human–Machine Mission Command

Bridging the gap between traditional CI and DCI will require the development of new skills in which commanders must be proficient. Following the transformation from an action-oriented (CI) to a governance-oriented (DCI) view, we identify four key skills as fundamental for a commander to deliver DCI. These include the AIAS capacity assessment, AIAS risk foresight, AIAS communication, and AIAS comprehension. Each skill follows a chain of skills for human–machine mission command (HMMC), including assessing the machine's capability, anticipating where the machine may fracture or fail, communicating the CI within appropriate bounds, and observing output behaviours to determine if the machine is executing as intended.

For homogeneous human teams, these skills are embedded as latent variables (for instance, shared understanding, trust, initiative, failure tolerance) inherent to the professional intuition and judgement of commanders experienced in the practice of command.³⁵ For heterogeneous human–machine teams, new competencies are necessary to interpret these distinct information signals. Table 1 summarises the key skills for each distinct logical task as phases of the commander–AIAS interaction.

Table 1: Proposed commander's skills for HMMC

Commander's skill for HMMC	Logical task	Lifecycle step	Function
AIAS capacity assessment	Assess machine capability	Assess	Evaluate what the machine can and cannot do, mapping performance envelopes to task demands
AIAS risk foresight	Anticipate where it may fail	Anticipate	Anticipate how the machine may begin to fracture or fail and where this may occur, identifying specific failure modes and environment stressors
AIAS communication	Communicate a bounded CI	Communicate	Deliver CI through the DCI, encoding governance bounds, outcomes, constraints, authorities and accountabilities
AIAS comprehension	Comprehend mission achievement	Comprehend	Interpret output behaviour to confirm execution aligns with the DCI

These skills are grounded in both military philosophical doctrine and HRI literature, providing a common foundation for proposing them. For the AIAS capacity assessment, military doctrine establishes the requirement for commanders to responsibly assess if a commander can execute the given task.³⁶ Previous human–automation research, adjacent to HRI, developed autonomy levels to aid humans in understanding machine capacity.³⁷ Function design when integrating humans and AIAS is a pertinent provision for dynamic situations requiring continuous reallocation of roles to agents,³⁸ with more recent work providing a formal representation of what an AIAS agent can do, defining the ontology that underpins a function space to evaluate capacity.³⁹ Without establishing a baseline representational structure, the commander and AIAS have no shared concept space to effectively map tasks to requirements. This is grounded through a risk-based framework, articulating what risks the AIAS introduces and how this links to the evaluation of delegation capacity.⁴⁰

AIAS risk foresight provides anticipatory reasoning about how and when the machine may begin to fracture and subsequently fail. Existing work has established that demonstrating why an agent is unreliable does not necessarily change trust but does provide a commander with awareness of the machine's limitations.⁴¹ Identifying novel machine-specific risks as failure modes that may impact the mission outcome is an essential task.⁴² A commander must have the capacity to understand, identify and consider these risks, a skill grounded presently only in homogeneous human teams. As commanders develop this skill over time, it will be important to train both over-reliance and under-reliance situations, with risk foresight providing awareness of the contextual capability and capacity bounds for an AIAS.⁴³

AIAS communication is the outbound-directional encoding of CI from the commander to the AIAS machine. Previous work provides our basis for this skill, establishing formal representation structures through which the CI can be encoded for machine consumption.⁴⁴ The requirement for bidirectional information flow represents CI through a structured representation rather than through a natural language paradigm alone. The rationale for this approach is established in the literature, with the limitations of direct communication highlighting that unstructured communication is insufficient for the delivery of a CI.⁴⁵ Recent work has framed machine education and communication as codependent pillars, suggesting that the communication skill is more than simply encoding intent—it is about ensuring that the skills to comprehend and negotiate intent are embedded in the pedagogical framework of an AIAS education and training architecture.⁴⁶

AIAS comprehension provides the inbound-directional interpretation of AIAS behaviour during execution—it is the bidirectional twin of AIAS communication. This skill focuses on the commander's ability to operationalise the tenets of transparency, including interpretability (can the commander understand what the AIAS is doing?), explainability (can the commander understand why?) and predictability (can the commander anticipate what will come next?) under real-world constraints.⁴⁷ The situation awareness-based agent transparency (SAT) model provides a basis of information levels that the machine must convey for the commander to exercise the skill, highlighting the bidirectional communication need for heterogeneous human–machine teams.⁴⁸ A primary outcome of AIAS comprehension is the commander being competent to manage DCI for heterogeneous human–AIAS teams.

While each skill is distinct, they necessarily have a directional dependency in practice. Figure 1 depicts the dependency structure of skills, leading to the commander attaining a competency in managing DCI for heterogeneous human–AIAS teams as the outcome of the commander's skills for HMMC. It also highlights the importance of the feedback loop as these skills continue to evolve as both the technology and environment continue to change.

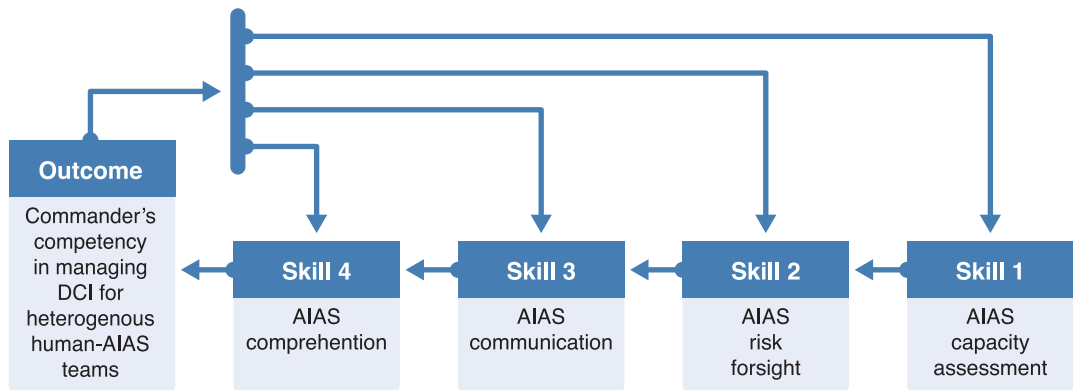


Figure 1: Dependency structure of commander's skills for HMMC

Looking Forward

The idea of DCI is not new, with established work on machine-interpretable representations of CI dating back to at least 2008 and the 13th International Command and Control Research and Technology Symposium (ICCRTS).⁴⁹ The term DCI has appeared in the literature as far back as 2015, originating from Defense Advanced Research Projects Agency and US Air Force Research Laboratory efforts,⁵⁰ with the use of the term in these settings extending traditional technical and doctrinal perspectives. Our governance-as-CI reframing, introduction of the four-skill commander framework for HMMC and explicit positioning of DCI as an enabling mechanism for MHC1 departs from the existing path.

To continue on this new path, an immediate first task is to formally specify the DCI for HMMC, including metrics for bidirectional comprehension, robustness, and stress testing to understand how elements of human–human team friction and fracture manifest in human–machine teams. Emerging discussions on test, evaluation and lifecycle assurance of military AI provides a foundational starting point for such metrics.⁵¹ Over the longer term, a well-developed DCI should inform practical evolutions of MC principles and requirements, specific to heterogeneous human–machine teams, generalising the key tenets for emerging heterogeneous teams.

In this paper, we have established an initial set of conditions for the introduction of a digital commander's intent (DCI), which will require new approaches to develop shared understanding between heterogeneous human–machine teams. This initial work articulates the problem space, identifies three key challenges for extending CI to mission command, and proposes a four-skill framework for commanders. Through this perspective, the DCI can be considered a generalisation of the commander's intent,

providing a structured, two-way artefact that a machine can act on and be governed through. The expansion beyond homogeneous human teams to heterogeneous human-machine teams necessitates a generalised approach to mission command (MC) that accounts for the heterogeneity inherent across culturally diverse biological and artificial agents. Realising DCI will necessitate the evolution of the existing MC paradigm to account for more artificial and heterogeneous agents, as well as the development of the necessary technical, policy and procedural interfaces and standards.

About the Authors

Lieutenant Colonel Dr Adam Hepworth is presently the Australian Army Liaison Officer to the United States Army Transformation and Training Command, based in Austin, Texas. Previously, Adam was Director of the Army Robotic and Autonomous Systems Implementation and Coordination Office (RICO). He is presently a Visiting Fellow at the University of New South Wales and an Advisory Board Member at the Responsible by Design Institute. Adam has formerly been an Expert Member on the Global Commission for Responsible AI in the Military Domain (GC REAIM), an expert advisor to the United Nations Institute for Disarmament Research (UNIDIR) on land warfare and autonomous systems, and an expert advisor to the International Committee of the Red Cross (ICRC) on Artificial Intelligence Decision Support Systems (AI-DSS). Adam holds a Bachelor of Science in Mathematics from the University of New South Wales, a Master of Logistics and Supply Chain Management from the University of South Australia, a Graduate Diploma in Scientific Computation and a Master of Science in Operations Research from the United States Naval Postgraduate School, and a Doctor of Philosophy in Computer Science from the University of New South Wales. He is a Graduate of the Australian Institute of Company Directors (GAICD).

Dr Hussein Abbass is a full professor with the School of Systems and Computing, University of New South Wales (UNSW), Canberra, served as the Deputy Head (People) (2021&22) School of Engineering & IT and the Executive Director (2006-2010) of the UNSW Defence and Security Applications Research Centre. Hussein was the National President (2016-2019) for the Australian Society for Operations Research, a member (2013-2015) of the Australian Research Council College of Experts and has been the Chair of the Corporate Governance Board for Opulence Institute of Higher Education since 2024. Hussein is a Fellow of IEEE, Australian Computer Society, UK Operational Research Society, and Australian Institute of Managers and Leaders. He is a Graduate Member of the Australian Institute of Company Directors. Hussein is a Distinguished Lecturer (2022-2024) for the IEEE Computational Intelligence Society (IEEE-CIS), Founding Editor-in-Chief (2020-2024) of the IEEE Transactions on Artificial Intelligence, Chair (2024-2025) of the IEEE Frank Rosenblatt Award Committee, vice-chair of the IEEE Standard 7018 for Security and Trustworthiness Requirements in Generative Pretrained Artificial Intelligence (AI) Models, and a member of the board for the Proceedings of IEEE.

ENDNOTES

- 1 International Committee of the Red Cross, 'ICRC Position on Autonomous Weapon Systems', *International Review of the Red Cross* 915 (2022), at: <https://international-review.icrc.org/articles/icrc-position-on-autonomous-weapon-systems-icrc-position-and-background-paper-915>.
- 2 Marta Bo, Laura Bruun and Vincent Boulanin, *Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS* (Stockholm: SIPRI, 2022).
- 3 Australian Defence Force, ADF Philosophical Doctrine 0 Series: ADF-P-0 *Command* (Canberra: Commonwealth of Australia, 2024).
- 4 Per M Gustavsson, Michael R Hieb, Philip Moore, Patric Eriksson, Lars Niklasson and Robbert van Geen, 'Machine Interpretable Representation of Commander's Intent', in *13th International Command and Control Research and Technology Symposium: C2 for Complex Endeavors* (ICCRTS, 2008). The term commandeer generalises the concept of command delegation, relaxing the subordinate constraint to not assume that delegation is to a lower rank—it could be at the same organisational level. Whereas the term subordinate assumes a hierarchal structure, 'commandeer' assumes a more general heterarchy, which may be more suitable in teaming contexts.
- 5 Adam J Hepworth, Zena Assaad, Alexander Wyatt and Hussein A Abbass, 'Meaningful Human Command: Towards a New Model for Military Human-Robot Interaction', in Simon Watson and Navinda Kottege (eds), *Comprehensive Robotics for Extreme and Challenging Environments* (Amsterdam: Elsevier, forthcoming). MHC1 is founded on the principles of mission command, as an organisational relationship to contextualise the operation of autonomous artificial cognitive agents within a human commander's intent. MHC1 is an evolution of meaningful human control, shifting from a control-centric model requiring direct supervision and intervention to a command-centric framing in which accountability is retained through bounded delegation rather than continuous oversight.
- 6 Australian Defence Force, ADF-P-0 *Command*.
- 7 Civil Air Patrol, *Commander's Intent: CAP Unit Commanders Course Student Guide* (Maxwell Air Force Base AL: Civil Air Patrol, n.d.).
- 8 UK Ministry of Defence, *Joint Doctrine Publication 0-01*, 6th edition (UK Ministry of Defence, 2022).
- 9 Canada Ministry of National Defence, *B-GL-300-001/FP-001 Land Operations* (Canada Ministry of National Defence, 2008).
- 10 Australian Defence Force, ADF-P-0 *Command*.
- 11 US Department of the Army, *ADP 6-0: Mission Command: Command and Control of Army Forces* (Washington DC: Headquarters, Department of the Army, 2019).
- 12 Australian Defence Force, ADF-P-0 *Command*.
- 13 Semantic compression, in this work, means conveying the maximum meaning (information signal) through the minimum words (information content) by relying on shared context.
- 14 Gustavsson et al., 'Machine Interpretable Representation'.
- 15 Anthony Duus, 'Paying the Mission Command Bill', *Australian Army Journal* 21, no. 3 (2025): 79–89; Dayton McCarthy, '“Anything but Simple and Clear Cut”: The Utility of Mission Command Within Domestic Security and Response Operations', *Australian Army Journal* 21, no. 3 (2025): 40–57; Sindre Sjøgren and Niklas Nilsson, 'Multinational Mission Command: From Paper to Practice in NATO', *Scandinavian Journal of Military Studies* 8, no. 1 (2025): 89–103; Hepworth et al., 'Meaningful Human Command'.
- 16 John Blaxland, 'Mission Command and the Australian Army (Mission Command Response)', *Australian Army Journal* 21, no. 3 (2025).
- 17 Blaxland, 'Mission Command and the Australian Army'.
- 18 Sjøgren and Nilsson, 'Multinational Mission Command'.
- 19 Blaxland, 'Mission Command and the Australian Army'.
- 20 Mica R Endsley, 'Toward a Theory of Situation Awareness in Dynamic Systems', *Human Factors* 37, no. 1 (1995): 32–64.

- 21 Australian Defence Force, *ADF-P-0 Command*, paras. 3.34, 6.11.
- 22 Ibid., para. 6.11.
- 23 Simon Stuart, *The Challenges to the Australian Army Profession: Address by the Chief of Army* (Canberra: AARC, 2024).
- 24 David S Alberts and Richard E Hayes, *Understanding Command and Control* (Washington DC: CCRP Publication Series, Department of Defense, 2006).
- 25 Ibid.
- 26 Frans PB Osinga, *Science, Strategy and War: The Strategic Theory of John Boyd* (London: Routledge, 2007). This assumption does not exclude connectionist architectures, where the full framework is encoded in a neural network without explicit non-overlapping modules for each phase (i.e. where the reasoning is spread across a neural network rather than separated into inspectable steps, so each phase is difficult to verify). This is a well-known assurance gap in mechanistic interpretability.
- 27 Hussein Abbass, Keeley Crockett, Jonathan Garibaldi, Alexander Gegov, Uzay Kaymak and Joao Miguel C Sousa, 'Editorial: From Explainable Artificial Intelligence (xAI) to Understandable Artificial Intelligence (uAI)', *IEEE Transactions on Artificial Intelligence* 5, no. 9 (2024): 4310–4314.
- 28 Hussein A Abbass, Eleni Petraki and Robert Hunjet, 'JSwarm: A Jingulu-Inspired Human-AI-Teaming Language for Context-Aware Swarm Guidance', *Frontiers in Physics* 10 (2022): 944064.
- 29 Australian Defence Force, *ADF-P-0 Command*.
- 30 Ibid.
- 31 US Department of the Army, *ADP 6-0*.
- 32 Hepworth et al., 'Meaningful Human Command'.
- 33 Bo, Bruun and Boulanin, *Retaining Human Responsibility*.
- 34 Australian Defence Force, *ADF-P-0 Command*.
- 35 Niklas Nilsson, 'Practicing Mission Command for Future Battlefield Challenges: The Case of the Swedish Army', *Defence Studies* 20, no. 4 (2020): 436–452; Blaxland, 'Mission Command and the Australian Army'.
- 36 Australian Defence Force, *ADF-P-0 Command*; US Department of the Army, *ADP 6-0*.
- 37 Raja Parasuraman, Thomas B Sheridan and Christopher D Wickens, 'A Model for Types and Levels of Human Interaction with Automation', *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 30, no. 3 (2000): 286–297.
- 38 Hussein A Abbass, 'Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust', *Cognitive Computation* 11, no. 2 (2019): 159–171.
- 39 Adam J Hepworth, Daniel P Baxter and Hussein A Abbass, 'Onto4MAT: A Swarm Shepherding Ontology for Generalized Multiagent Teaming', *IEEE Access* 10 (2022): 59843–59861.
- 40 Zena Assaad, 'A Risk-Based Trust Framework for Assuring the Humans in Human-Machine Teaming', in *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems* (New York: Association for Computing Machinery, 2024), pp. 1–9.
- 41 Aya Hussein, Sondoss Elsayah and Hussein A Abbass, 'The Reliability and Transparency Bases of Trust in Human-Swarm Interaction: Principles and Implications', *Ergonomics* 63, no. 9 (2020): 1116–1132.
- 42 Mica R Endsley, 'From Here to Autonomy: Lessons Learned from Human-Automation Research', *Human Factors* 59, no. 1 (2017): 5–27.
- 43 John D Lee and Katrina A See, 'Trust in Automation: Designing for Appropriate Reliance', *Human Factors* 46, no. 1 (2004): 50–80.
- 44 Hepworth, Baxter and Abbass, 'Onto4MAT'.
- 45 Maarten Schadd, Anne Merel Sternheim, Romy Blankendaal, Martin van der Kaaij and Olaf Visker, 'How a Machine Can Understand the Command Intent', *The Journal of Defense Modeling and Simulation* 22, no. 1 (2025): 41–58.
- 46 Hussein Abbass, Eleni Petraki, Aya Hussein, Finlay McCall and Sondoss Elsayah, 'A Model of Symbiomemesis: Machine Education and Communication as Pillars for Human-Autonomy Symbiosis', *Philosophical Transactions of the Royal Society A* 379, no. 2207 (2021): 20200364.

- 47 Adam J Hepworth, Daniel P Baxter, Aya Hussein, Kate J Yaxley, Essam Debie and Hussein A Abbass, 'Human-Swarm-Teaming Transparency and Trust Architecture', *IEEE/CAA Journal of Automatica Sinica* 8, no. 7 (2021): 1281–1295.
- 48 Jessie YC Chen, Kimberly Procci, Michael Boyce, Julia Wright, Andre Garcia and Michael Barnes, *Situation Awareness-Based Agent Transparency*, Report No. ARL-TR-6905 (Aberdeen Proving Ground MD: US Army Research Laboratory, 2014); Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright and Michael Barnes, 'Situation Awareness-Based Agent Transparency and Human-Autonomy Teaming Effectiveness', *Theoretical Issues in Ergonomics Science* 19, no. 3 (2018): 259–582.
- 49 Gustavsson et al., 'Machine Interpretable Representation'.
- 50 Stephen Lee-Urban, Ethan Trehwhitt, Ian Bieder, Joel Odom, Timothy Boone and Elizabeth Whitaker, 'CORA: A Flexible Hybrid Approach to Building Cognitive Systems', in *Proceedings of the Third Annual Conference on Advances in Cognitive Systems*, Poster Collection (Cognitive Systems Foundation, 2015).
- 51 David Helmer, Michael Boardman, S Kate Conroy, Adam J Hepworth and Manoj Harjani, 'Human-Centred Test and Evaluation of Military AI', *arXiv* (2024), at: <https://doi.org/10.48550/arXiv.2412.01978>; Zena Assaad and Adam Hepworth, *A Systems Engineering Lifecycle Approach to Responsible AI* (The Hague: The Hague Centre for Strategic Studies, 2025).

/ GOVERNING JOINT INNOVATION: LEGAL REVIEW AND EMERGING TECHNOLOGIES UNDER AUKUS PILLAR 2

Natalia Jevglevskaja and Renato Wolf

Introduction

Although public attention since the September 2021 announcement of the AUKUS partnership¹ between Australia, the United Kingdom (UK) and the United States has largely (US) centred on nuclear-powered submarines,² there is broad agreement that the development and sharing of advanced capabilities under Pillar 2 will be equally central to the partnership's success.³ Driven in significant part by China's rapid and unprecedented military modernisation – including in areas such as artificial intelligence, hypersonic systems and electronic warfare – the Indo-Pacific is experiencing accelerated technological change in the defence domain. Against this backdrop, it is essential that the AUKUS partners retain a technological edge in order to sustain credible deterrence and contribute to regional stability.⁴ Pillar 2 responds to this imperative by establishing a trilateral capability accelerator that targets technologies of immediate strategic relevance, linking government, industry and research communities across the three States to enhance interoperability and promote collaborative innovation.⁵

The apparently intensifying pace of advanced capability development under Pillar 2 creates a parallel imperative: ensuring that relevant governance frameworks – including legal review processes – evolve at a comparable pace. Certain Pillar 2 technologies – most notably AI-enabled systems – are not static artefacts but adaptive systems whose behaviour may change over time through software updates, data inputs and integration with other (weapon) systems and platforms. This dynamism has already strained national legal review frameworks, which were largely designed for stable, predictable technologies, prompting ongoing efforts to recalibrate review methodologies.⁶ Other Pillar 2 technologies – such as quantum capabilities and advanced electronic warfare systems – are likely to test these frameworks in comparable, if not more profound, ways, further exposing the limits of existing approaches.

Crucially, Pillar 2 requires policymakers and legal review authorities to think one step ahead. If capabilities developed under Pillar 2 are intended to be designed, tested and deployed jointly, it is hard to see how this can be meaningfully achieved without triangulating legal considerations relevant for each AUKUS partner. In fact, this has prompted at least one analyst to call for ‘multinational’ – that is, AUKUS-specific – legal review processes.⁷ While such proposals have intuitive appeal, the more difficult question is what form such a review mechanism would take, how it could operate in practice as a single process, and whether it would be viable in the first place given that the three AUKUS partners might be bound by different international law obligations in relation to capability development and also operate under distinct – and at times more restrictive – national legal and policy frameworks.

Despite its significance, the implications of AUKUS Pillar 2 for the legal and policy requirements to conduct legal review of military capabilities⁸ have received little, if any, sustained attention in the expert literature. Addressing this gap, the paper critically examines the frictions arising from the divergent approaches to legal reviews across the three AUKUS partners and explores the viability and limits of possible mitigation measures. Specifically, we first discuss the feasibility of **a joint legal review mechanism** where the capabilities developed jointly are also reviewed jointly. We then examine an alternative measure – **a trilateral legal review coordination mechanism** – where the national review processes can be coordinated, results shared and problems addressed. The aim of proposed mitigation measures is to support the objectives of Pillar 2 and identify key legal, policy and practical challenges that such mechanisms would need to address. For the Australian Army, the issue is of particular relevance as many emerging capabilities – especially AI-enabled systems – may be developed in the framework of Pillar 2, and issues with legal review may thus particularly affect the Australian Army. The purpose of the paper is therefore to draw attention to an issue of growing importance for Australia and its AUKUS partners and to stimulate more systematic discussion of how legal review processes should evolve in the context of jointly developed emerging technologies.

In what follows, **the paper** undertakes a stocktaking exercise of AUKUS Pillar II, assessing the extent to which its ambitious commitments to advanced capability development and industrial integration have translated into tangible outcomes. **The paper** outlines the national legal review processes in Australia, the UK and the US and identifies key areas of convergence and divergence. **Later in the paper** we examine the implications of these similarities and differences for the feasibility of a joint legal review mechanism, taking into account the broader practical challenges currently confronting Pillar 2. In view of significant legal and practical hurdles to the establishment and maintenance of a joint mechanism, **the paper** outlines key aspects of a more likely arrangement that AUKUS partners might pursue – a trilateral legal review coordination mechanism.

AUKUS Pillar 2. Nearly Five Years In: Where Are We Now?

The creation of AUKUS was driven by the recognition ‘that the Indo-Pacific is entering a period of unprecedented strategic volatility’.⁹ While Pillar 1 of the partnership seeks to enable the acquisition of nuclear powered submarines by Australia, Pillar 2 comprises ‘a suite of cooperative activities’ through which the United States, the United Kingdom, and Australia seek to ‘develop and field “advanced capabilities”’.¹⁰

Initially, the AUKUS leaders announced four areas of cooperation in critical technologies, namely ‘cyber capabilities, artificial intelligence, quantum technologies, and additional undersea capabilities’.¹¹ However, not long thereafter, in April 2022, these were expanded to include two further *technology-focused* areas (hypersonic and counter-hypersonic capabilities and electronic warfare capabilities) and two *functional* areas, namely innovation and information sharing.¹² The three States have been coordinating Pillar 2 activities through various mechanisms, including working groups (also known as ‘workstreams’).¹³

The technological breadth of the undertaking is extraordinary. Unlike Pillar 1, centred on nuclear-powered submarines, Pillar II encompasses a broad, evolving suite of advanced and emerging technologies. Each of the listed technological domains represents a complex ecosystem involving software, hardware, data infrastructure, testing environments, and operational doctrine. While some capabilities sit within specialised domains, such as undersea systems and hypersonic technologies, others – most notably AI, quantum technologies, and cyber capabilities – are cross-cutting in nature. With applications extending across military and civilian sectors, they are likely to define the trajectory of warfare and security at national, regional, and global levels well into the future. Sustained joint technological progress will therefore be essential for the AUKUS partners to outpace strategic competitors in this field, notably key Indo-Pacific actors such as China and India.¹⁴

The timing of Pillar 2 appears particularly opportune. Analysis by the Australian Strategic Policy Institute indicates that China currently leads in high-impact research in 19 of 23 critical technologies, with a commanding position in areas such as hypersonics, electronic warfare and key undersea capabilities.¹⁵ Still, the picture changes when the *collective* technological capacity of the AUKUS partners is considered. In several strategically important fields – including ‘autonomous systems operation technology, advanced robotics, adversarial AI reverse-engineering and protective cyber’ – the combined strength of Australia, the UK and the US places them at the forefront globally.¹⁶ These developments take place against the backdrop of China’s innovation-driven development strategy, launched in 2016,¹⁷ which aims to establish the country as a global leader in science and technology by 2050.¹⁸ Pillar 2 thus seeks to steer its members’ efforts in counterbalancing China’s rivalry for technological hegemony.

Despite the central role attributed to Pillar 2 in the partnership's long-term strategic vision, publicly available information about its practical implementation remains limited. As commentators have observed, access to the AUKUS material remains scarce not only because programs such as the SSN-AUKUS submarines will take decades to deliver, but also because both Pillar 1 and Pillar 2 involve highly sensitive technologies whose disclosure is tightly controlled.¹⁹ That said, calls for the creation of a 'Pillar 3' – variously proposed to address issues such as critical minerals²⁰ or space capabilities²¹ – have been voiced since 2024, even as the operational contours of Pillar 2 itself remain only partially understood.

The enabling architecture for Pillar 2 is still evolving. Important legislative and regulatory steps have been taken, including recent adjustments to export control frameworks in all three States.²² Yet assessments of progress remain mixed. While some commentators defend the significance of these reforms,²³ others point to persistent structural obstacles, particularly in areas such as the US Foreign Military Sales system, which is regarded as slow, rigid and ill-suited to collaborative development of rapidly evolving technologies such as AI or quantum systems.²⁴ More broadly, analysts note that Pillar 2 has struggled to generate tangible outcomes and faces pressure to demonstrate clearer priorities.²⁵ Nevertheless, even if the scope of Pillar 2 is eventually narrowed or refocused, its central objective – enhanced interoperability through the joint development of advanced capabilities – will remain fundamental. This enduring emphasis on collaborative capability development reinforces the importance of addressing the legal and governance questions examined in this paper, particularly if Pillar 2 expands in the future to include additional partners with differently structured legal review mechanisms, such as New Zealand or Canada, or partners for which no comparable mechanisms are publicly known, such as Japan or South Korea.²⁶

National Legal Reviews: Points of Convergence and Divergence

A central issue for the AUKUS partners is whether and how the ambitions of Pillar 2 can be pursued consistently with their obligations under international law and other applicable domestic law and policy commitments. Although the expert literature suggests that international law places relatively few restrictions on the inter-State sharing of advanced technologies of the kind envisaged under Pillar 2,²⁷ the joint design and development of military capabilities engage a range of international legal obligations. Among others, all three AUKUS States must ensure compliance with the law of armed conflict (LOAC), which applies to their cooperative activities in this area.

One of the core principles of LOAC, set out in Article 35(1) of Additional Protocol I to the Geneva Conventions (AP I / Protocol), is that the right of belligerents to choose weapons, means, and methods of warfare is not unlimited.²⁸ Article 36 of the Protocol gives practical

effect to this principle by requiring States, prior to the development or acquisition of a military capability, to assess whether it can be used consistently with international law.²⁹ Described commonly as the duty to conduct ‘weapons reviews’ or ‘legal reviews’ (for the further: legal reviews) this requirement may be tightened by States to incorporate more restrictive obligations under domestic law or policy.³⁰

In what follows, we briefly outline domestic legal review regimes in the three States. These regimes display both commonalities and points of divergence in the framing and conduct of legal reviews (section C2). While areas of convergence may facilitate the development of a joint approach under AUKUS, differences in scope, institutional design, or review practice may signal potential pressure points and practical hurdles for any such future cooperation.

Legal Reviews: Implementation in Australia, the United Kingdom and the United States

Australia and the UK are both parties to AP I and each conducts legal reviews pursuant to their obligations under Article 36 of the Protocol. The United States, by contrast, has never ratified this instrument, although it is a signatory. Notably, the United States’ practice of reviewing weapons predates the Protocol’s adoption in 1977, with documented examples of such reviews as early as 1918, undertaken as part of an informal review practice.³¹ The US conducts weapons reviews as a matter of policy treating them as a best practice for implementing customary and treaty-based constraints on weapons use.³²

Australia

Australia formally assumed the obligation to conduct legal reviews in June 1991 upon ratifying AP I. For more than a decade, legal reviews were performed on an *ad hoc* basis, that is without a structured institutional framework. This changed in June 2005, when the Australian Defence Force (ADF) formalised the review process through a Defence Instruction which established a consistent procedure within the ADF by laying out roles and responsibilities for the conduct of legal reviews.³³

The directive was subsequently revised on several occasions. In May 2020, the process was updated to ensure closer alignment with Defence’s capability acquisition lifecycle. As part of this revision, legal advice under Article 36 became integrated into high-level decisions, including early-stage consideration of whether proposed capabilities should proceed to procurement.³⁴

Another series of reviews were undertaken between 2022 and 2023 and resulted in the adoption of the *Guide to the Legal Review of New Weapons, Means or Methods of Warfare (ADF Guide)*, which remains the legal basis for conducting reviews in Australia at the time of writing.³⁵ The Guide is organised into three parts. Part 1 sets out key

definitions, procedures, and responsibilities. Overall responsibility for Article 36 reviews rests with the Director-General of the Military Legal Service, while assessments are conducted in practice by the Directorate of Operations and International Law (DOIL) at the request of authorised Defence entities. Capability managers initiating a review are required to complete an online legal review report form and provide supporting technical documentation.³⁶ DOIL then conducts the assessment based on the submitted material. Each review is undertaken by two legal officers: a primary reviewer and a senior reviewer exercising delegated authority. The completed assessment is formally counter-signed by the requester. Part 2 of the Guide outlines Australia's established interpretation of the scope of Article 36, and Part 3 addresses additional considerations relevant to the assessment of autonomous weapon systems (AWS).

United Kingdom

The UK ratified AP I in January 1998. Although legal reviews of military capabilities had been conducted before that date on an informal basis, ratification prompted the establishment of a structured process for legal reviews. Since then, the process has significantly evolved in response to developments in military technology and procurement practice.³⁷

Today, legal reviews are conducted on behalf of the Ministry of Defence by a tri-service team of military legal advisers within Defence Futures and Force Design (F&FD), previously known as Defence Futures (DF) and, before July 2024, the Development, Concepts and Doctrine Centre (DCDC).³⁸ Having one legal officer from each Service ensures that all Services remain actively engaged in the discharge of the UK's Article 36 obligations.

The review process is governed by a Joint Service Publication which is not in the public domain.³⁹ The Publication sets out the applicable procedures and includes a template that capability owners must complete in order to initiate a review.⁴⁰ Once the request is received, the reviewing authority develops an assessment plan and requests relevant information at the earliest opportunity. Since the information required varies from one capability to the next, the team consults with appropriate subject-matter experts to develop the necessary technical and evidentiary basis for the assessment. Final advice is typically delivered before the capability is introduced into the UK Armed Forces' force structure with interim advice governing capability development where relevant.⁴¹

United States

While, as mentioned, the US is not formally bound by the AP I, legal reviews of weapons have been a formal policy requirement since 1974.⁴² The responsibility for conducting these reviews is closely linked to acquisition authority within the Department of Defense (DoD). The DoD component responsible for acquiring or developing a weapon system

is also responsible for conducting its legal review. Accordingly, military services – the Army, Navy, and Air Force, as well as certain Combatant Commands – including Cyber Command and Special Operations Command – review the legality of the weapons they acquire or develop.⁴³

While procedural steps vary slightly among the services, there are no substantial differences in the standards applied.⁴⁴ Each review results in a written legal opinion. Draft opinions are typically circulated to the relevant Project Office to confirm technical specifications and ensure appropriate classification markings. The draft is usually shared with other services for concurrence; however, such consultation does not remove the requirement for each service to conduct its own review.⁴⁵

Points of Convergence and Divergence

It may appear intuitive that capabilities developed jointly under AUKUS Pillar 2 should also be reviewed jointly. Michael Sharpe, Chief Executive Officer of the AUKUS Forum,⁴⁶ has observed that he likes to think of AUKUS as ‘one big factory across three nations.’⁴⁷ From a legal review perspective, such an integrated model of capability development may equally call for a single joint legal review across all three jurisdictions. But is this feasible? This section focuses on the points of convergence and divergence in the practice of legal reviews, in order to assess whether the future establishment and operation of ‘a joint legal review mechanism under AUKUS Pillar 2’ may be facilitated or constrained.

Points of Convergence

To begin with, in all three AUKUS States, **the authority to conduct legal reviews is vested within their respective departments of defence** and carried out by subordinate units within those institutions. The UK’s Defence Futures and Force Design (F&FD) is an organisational unit within the Ministry of Defence (responsible inter alia for future capability planning and concept development) and review authorities both in Australia and the US (DOIL and individual services⁴⁸) equally sit within respective Departments of Defence. We believe that the embeddedness of legal reviews within the defence establishment itself, suggests a broadly similar institutional setting, culture and mode of operation – an aspect likely to facilitate the development of a joint review mechanism.

Contrast this with the Swedish legal review model, where the Delegation for International Law Monitoring of Arms Projects, established by government ordinance in 1974, operates as an independent and multidisciplinary body formally separate from both the government and the armed forces.⁴⁹ It appears that the institutional setting in the AUKUS States is likely to suggest more comparable working practices and professional outlooks, as opposed to approaches where the legal review function is assigned to an independent body outside the defence establishment. An external review authority may operate with greater structural independence from capability development processes and also follow different procedural preferences and rhythms.

Second, in all three States, **legal reviews function as a component within a broader governance architecture designed to ensure that new capabilities can be employed consistently with international law**, in particular IHL.⁵⁰ Within this broader governance structure, the legal review operates alongside a range of other legal, technical, and operational processes that collectively ensure compliance with applicable law. It plays a **specific preventive role**: it assesses whether a military capability is capable of being used lawfully when it is introduced into service.⁵¹ Ensuring that the weapon is indeed used lawfully in practice depends on the wider institutional framework governing military operations. Australia explicitly articulated this position in 2019 in its ‘system of control’ governing the military use of force.⁵² Accordingly, the legal review process forms *one* (albeit important) component of a layered framework covering ‘all aspects of a weapon system from design through to engagement,’⁵³ including weapon design, testing, evaluation and review, training and certification, rules of engagement, and command oversight.

The UK and the US positions are similar. Officials responsible for legal reviews in the UK emphasise that they constitute an important component of the UK’s international law compliance processes but are only one element within a broader defence framework aimed at ensuring the lawful development and deployment of military capabilities.⁵⁴ The US equally lists legal reviews as one of the requirements within the broader framework of compliance with the laws of war.⁵⁵

The fact that all three States conceptualise legal reviews as one element within a broader system of the capability development and use reflects a shared governance philosophy. This common understanding of the role of legal reviews within capability acquisition is, in our opinion, likely to facilitate coordination between the three systems, even if the precise institutional and governance arrangements differ.

Third, legal review processes in the three AUKUS States have **evolved** continuously in response to both technological developments and changes in defence capability procurement systems. In particular, all three States in the past decade turned their attention to the challenges posed by AI capabilities and increasingly complex processes of human–machine interaction, reflecting a shared recognition that traditional legal review procedures require recalibration to address software-driven systems whose operational behaviour may change over time. The three States continue to explore ways of integrating legal reviews more closely into capability acquisition cycles so that legal advice can inform design, testing, and development decisions at earlier stages.⁵⁶ This convergence in regulatory focus, timing, and appreciation of the challenges ahead creates favourable conditions for the development of a joint review mechanism, as the participating States share a common understanding of the legal and operational implications of emerging technologies and the role of legal reviews in addressing them.

Points of Divergence

Whilst certain aspects of the national legal review processes suggest feasibility of a joint review mechanism, important points of divergence between national systems may pull in the opposite direction.

The scope of legal reviews

To begin with, important differences arise in the way Australia and the UK define **the scope of legal reviews** compared to the US. Ultimately, the US's narrower approach creates uncertainty as to what would be reviewed jointly and risks excluding capabilities that Australia and the UK would otherwise review. Our reasoning is as follows.

Both States explicitly frame the scope of their legal review obligation as covering 'weapons' and 'means of warfare', as per Article 36 AP I. Australia's Guide defines 'a weapon' broadly as any 'arms, munition, device, whether tangible or intangible, designed or intended to be used to cause: (1) injury to, or death of, persons; or (2) damage to, destruction of or neutralisation of objects' and explains that 'means of warfare' include 'weapons as well as weapon systems, which includes all related equipment, materials, and delivery mechanisms necessary for a weapon to function as intended'.⁵⁷ The UK practice likewise adopts an expansive interpretation of 'means of warfare.' Earlier UK guidance indicated that Article 36 reviews encompass 'weapons in the broader sense of that word' as well as 'sighting equipment, laser designators, target acquisition equipment, data links and software used for processing targeting data', and even capabilities such as tactical radios or navigation equipment may be submitted for review before a determination is made that a formal assessment is unnecessary.⁵⁸

By contrast, the US formally focuses on the review of 'weapons' and 'weapon systems.' Definitions across the US services generally describe weapons as devices intended to injure or kill persons or damage or destroy property.⁵⁹ This terminology may appear narrower because US approach does not explicitly refer to 'means of warfare.' However, some US definitions of weapons and weapon systems are so expansive – covering a wide range of components and equipment required for their operation – that, in practice, the scope of review may capture much, if not all, of what other two States describe as 'means of warfare'. Consider, for example, the US Air Force definition of a 'weapon system' which includes 'a combination of one or more weapons with *all related equipment, materials, services, personnel, and means of delivery and deployment* (if applicable) required for self-sufficiency'.⁶⁰

The divergence is, however, more profound for **launch or delivery platforms**. The US services explicitly exclude such platforms – for example, ships, aircraft, or armoured vehicles – from the definition of weapons subject to review.⁶¹ By contrast, the UK and Australian approaches do not appear to exclude such systems a priori. Indeed, Australia's

definition of ‘means of warfare’ refers to the ‘delivery mechanisms necessary for a weapon to function as intended’, suggesting that the scope of review may extend more readily to the platforms that enable the use of force. Likewise, nothing in the UK’s broad conceptualisation of weapons and means appears, at least *prima facie*, to exclude such platforms from consideration.

Another divergence arises with respect to ‘methods of warfare.’ Australia defines methods broadly as ‘the manner in which weapons and certain capabilities are used in warfare to harm, destroy, or neutralise the enemy’,⁶² and UK practice likewise considers methods to extend beyond ‘normal’ or ‘anticipated’ uses of weapons to ‘ways in which warfare is conducted’.⁶³ US services, in comparison, generally assess the ‘concept of employment’ *aka* ‘intended use’ of a weapon, rather than reviewing methods of warfare as such.⁶⁴ At first glance, this approach appears narrower. In the US practice, however, broader methods of warfare (including, for example, tactics, techniques and procedures) are not ignored but may instead be addressed through separate policy, doctrinal, and operational oversight processes outside the formal legal review regime.⁶⁵ In contrast, in the UK, the inclusion of a wider range of methods in the legal review framework reflects the institutional remit of F&FD, whose focus on strategic foresight, concepts, capability strategy, and force design analysis necessarily engages with a broader issue of methods employed in warfighting.

The institutional separation reflected in the US treatment of broader methods of warfare – compared to the more integrated approaches of Australia and the UK – may complicate the design of a joint AUKUS review mechanism and weaken its prospects. At a minimum, it introduces uncertainty as to the appropriate scope of any joint review and the timing of assessments within the capability lifecycle. Aligning these approaches would require agreement on whether such methods should be incorporated into a single review process or continue to be addressed through parallel mechanisms, thereby adding layers of coordination and potential friction.

Standard of review: Superfluous Injury and Unnecessary Suffering

Another point of divergence concerns how the three States apply the rule prohibiting the infliction of superfluous injury or unnecessary suffering (SI/US) in their review processes.⁶⁶ While all three States largely follow the same two-step methodology – examining first compliance with *specific* prohibitions or restrictions under treaty or customary IHL and then with *general* prohibitions or restrictions under treaty and custom, including whether the weapon is capable of being used in accordance with the principle of distinction and whether it causes SI/US – differences arise in how the latter standard is interpreted and applied.

There is broad agreement on the substantive purpose of the rule. As the International Court of Justice explained in the *Nuclear Weapons Advisory Opinion*, the prohibition concerns weapons that cause ‘harm greater than that unavoidable to achieve legitimate military objectives’.⁶⁷ The rule therefore requires a balancing of the humanitarian harm caused by the weapon against the military advantage expected from its use. The central (and contentious) issue for present purposes is what gives rise to the harm – whether it is the inherent *nature* of the weapon or the *intention of its designer*.

Australia and the UK apply the standard codified in Article 35(2) of AP I, which prohibits the use of weapons ‘of a nature to cause’ SI/US. Under this formulation, the legality of a weapon depends primarily on its inherent characteristics and effects.

Although strong arguments suggest that the customary IHL test is that ‘of a nature’,⁶⁸ the US has long taken the position that the correct formulation of the rule is that weapons must not be ‘*calculated to cause superfluous injury*’. This approach, first adopted in the 1956 *U.S. Army Field Manual on Land Warfare*,⁶⁹ reinforced since by prominent US commentators⁷⁰ and currently reaffirmed in the *U.S. Department of Defense Law of War Manual*,⁷¹ appears to focus on the *purpose or design intention* behind the weapon. In the US view, the phrase ‘calculated to cause’ more accurately reflects the intent of the original French text of the 1907 Hague Regulations.⁷²

The difference in tests applied is not merely theoretical. Consider **an AI-enabled automated sentry weapon designed to neutralise approaching enemy personnel**.

The system uses computer vision to identify human movement and is programmed to fire repeated low-calibre rounds at a detected combatant until the algorithm determines that the person is no longer mobile. The design objective is operationally framed as ensuring reliable neutralisation in high-tempo environments where human operators may not be able to intervene quickly.

Under the US ‘calculated to cause’ SI/US formulation, the relevant question is whether the weapon was designed with the purpose of increasing suffering beyond what is required by military necessity. In this case, it could be plausibly argued that the system is intended simply to ensure neutralisation of the target, not to increase suffering. The repeated firing pattern arises from the algorithmic logic used to confirm incapacitation, not from an intention to aggravate injury.

However, under the ‘of a nature to cause’ SI/US standard applied by Australia and the UK pursuant to Article 35(2) AP I, the analysis would focus on the predictable effects of the weapon in operation.⁷³ A system that systematically continues to fire at already incapacitated personnel until the algorithm registers immobilisation could foreseeably produce multiple and unnecessary wounds beyond those required to neutralise the combatant. Because other anti-personnel weapons – for example, those used by human

sentry personnel – can achieve the same military purpose (incapacitation) without predictably causing such cumulative injury, the system could be considered ‘of a nature’ to cause unnecessary suffering.

Given these differences in application, any joint legal review mechanism would, at a minimum, need to address explicitly the competing formulations of the rule and determine how to proceed where those differences lead to divergent outcomes – such that the same capability may be permissible for the US but not for the UK and Australia.

Standard of review: Environmental damage

A further point of divergence concerns the protection of the natural environment in armed conflict. Australia and the UK are bound by Art. 35(3) AP I, which prohibits the employment of methods or means of warfare that are intended, or may be expected, to cause ‘widespread, long-term and severe damage to the natural environment’. Although there is now sufficient State practice to support the proposition that this rule has crystallised to customary international law,⁷⁴ the US has positioned itself as a persistent objector to it and does not include the rule in its standard of legal review.⁷⁵

This divergence in approaches suggests that a capability developed jointly under AUKUS Pillar 2 might be assessed against different environmental standards by the participating States. Consider a long-endurance autonomous underwater system designed to deploy large numbers of smart naval mines in sensitive maritime environments or an AI-enabled autonomous area-denial systems designed to ignite large-scale vegetation fires in order to restrict enemy movement. If the expected environmental effects of such systems included a widespread, long-term and severe disruption of affected marine ecosystems in the former case and a widespread, long-term and severe degradation of vegetation, soil and wildlife habitats in the latter, the systems would be impermissible under the standard applied by the UK and Australia but likely permissible under that applied by the US.

This issue is not one of simple arithmetic where the position of two States would automatically prevail over that of one. The mere existence of a joint arrangement would not, without more, require the US to revise its settled interpretations of international law or assume obligations it has consistently declined to recognise. In practice, however, this could create a structural difficulty for jointly developed capabilities under Pillar 2, as joint funds might be invested in a capability that ultimately can be lawfully employed by only one partner.

One possible response would be for Australia and the UK to insist, at an early stage of capability development, that systems likely to meet the environmental damage threshold should not be pursued within the Pillar 2 framework. However, apart from the US expected objections to that as a matter of law, such an approach also raises

practical and political concerns. Predicting environmental effects decades in advance – particularly for complex, software-enabled systems – may be difficult, and imposing *ex ante* constraints could limit technological experimentation or innovation. The feasibility of such an approach would therefore depend on whether the partners are willing to accept stricter common standards for jointly developed systems or, alternatively, to tolerate the possibility that certain capabilities will ultimately be usable by only some or one of them.

Standard of review: international law vs additional legal considerations

Differences in States' international legal obligations, together with the inclusion of domestic law (and policy) in national legal review processes could pose yet another significant hurdle for the establishment of a joint legal review mechanism. In the US, reviews conducted by the military services explicitly assess whether a weapon / weapon system complies not only with international law but also with domestic law and policy.⁷⁶ Likewise, although the primary focus of Australia's Article 36 review is compliance with international law, the Australian framework recognises that issues arising under domestic law may surface and require attention during the review process.⁷⁷ The UK similarly frames legal reviews primarily in terms of international law obligations, although in practice it would be difficult for reviewing authorities to disregard relevant domestic legal constraints.⁷⁸

This issue is significant because domestic legal frameworks differ substantially between States and may impose additional or distinct constraints on the development, acquisition or use of military capabilities. While a State cannot invoke its domestic law to avoid its obligations under international law, nor require another State to comply with its national law,⁷⁹ these differences nonetheless complicate the operation of a joint review mechanism. The concern here is that more restrictive domestic standards applied by one partner could, in effect, be imposed on others through such a mechanism, thereby bypassing ordinary domestic legislative or regulatory processes. However, even if the AUKUS partners were to amend their domestic legal and regulatory frameworks to accommodate restrictions applied by other partners within a joint review mechanism – that is by creating an explicit legal basis for incorporating more *restrictive* standards applied by any of the AUKUS partners – it is difficult to imagine that the three AUKUS States would go down that path. After all, a mechanism designed to produce a single determination of legality acceptable to all three partners could prohibit or restrict a capability that would otherwise be lawful under the applicable rules of international law. In circumstances where a State would have been free to develop and employ the same capability if acting independently, such an outcome would be difficult to justify.

Any joint mechanism would therefore need to clearly define how to address situations in which participating States are bound by differing international or domestic legal obligations, or where they adopt divergent interpretations of shared legal standards.

In that regard, a number of Workshop participants expressed concern that a joint review mechanism could reduce or dilute the applicable legal standard, ultimately rendering it little more than a fig leaf. They pointed to what they termed the ‘least common denominator’ problem, which could manifest in two ways.

First, if one categorises the relevant law into four groups – Category A (international and domestic law binding on Australia), Category B (international and domestic law binding on the UK), Category C (international and domestic law binding on the US), and Category D (treaty and customary IHL rules binding on all three) – it might initially seem that cooperation under AUKUS Pillar II could extend only to capabilities falling within Category D, and not to those affected by Categories A, B, or C.

Consider, for example, the Convention on Cluster Munitions.⁸⁰ the US is not a party, whereas Australia and the United Kingdom are. On this logic, Category D would exclude obligations arising under that Convention. The concern, put differently, is that the collective baseline would reflect only the lowest common set of shared obligations, thereby reducing the level of scrutiny applied to a capability under development. In the example above, this raises the risk that, within a joint legal review framework, obligations under the Convention could be sidelined because they do not bind all three partners.

Another way to frame the ‘least common denominator’ problem is as a selective, issue-by-issue approach in which, for each legally relevant aspect of a capability, the most permissive standard among the three States is applied. Thus, if the US adopts the least restrictive interpretation of unnecessary suffering (see our discussion in section 2 ii) b) above) that standard would govern that element of the joint review; if Australia applies the least restrictive approach to indiscriminate weapons, its standard would govern that element, and so on. The likely consequence is that the joint review would deem a capability lawful for all three States, even though particular features of that capability would fail to meet the stricter standards each State applies in its own system. In effect, the process would aggregate permissive elements across jurisdictions, resulting in a systematic lowering of the overall level of legal scrutiny.

We suggest, however, that this is not how practice is likely to operate. Each of the three States has thus far conscientiously reviewed weapons against the international and domestic legal obligations binding upon it. It is difficult to imagine that they would disregard applicable law merely because a capability is being developed jointly under AUKUS Pillar II. Rather, reviewing authorities in each State are likely to identify the international and domestic rules applicable to them and to specify any developmental, acquisition, or operational limitations required by those obligations. For example, where a capability engages obligations binding on Australia or the UK but not the US, those States may approve participation subject to conditions – such as restrictions on design features, targeting parameters, or permissible operational use – while the US conducts its

review against its own set of legal obligations and interpretations. In this way, divergence in applicable law is managed through nationally tailored caveats rather than by lowering the overall standard of review. It is for these reasons that we consider a Trilateral Coordination Mechanism to be the most appropriate framework for managing legal review processes in respect of AUKUS Pillar II capabilities (see section E below).

One final remark is appropriate here. A joint legal review that genuinely answers the question of whether a capability may be used lawfully by each State would require precisely the opposite logic to that feared under a ‘least common denominator’ approach. In theory, it cannot be discounted, that rather than defaulting to the most permissive rule, a joint legal review would apply, for each legally relevant aspect, the most restrictive standard among the participating States. If Australia adopts the most stringent position on unnecessary suffering, that standard would need to govern that element of the assessment; if the US applies the strictest approach to indiscriminate weapons, its standard would control that element, and so on. The practical consequence is that any capability capable of passing such a composite review would, by design, reflect the highest level of constraint across the three systems. The difficulty of such an approach is, however, that each state will likely have restrictions on the use of a given capability they would not otherwise have if they only applied their own legal standards. While the approach may be lawful, it thus likely limits the overall military utility of a given capability..

One Joint Legal Review: Meaningful and Feasible or Rather Not?

As the preceding discussion illustrates, in some respects, the ambition of ‘one joint legal review’ is supported by existing convergence between the national practices of legal reviews. As shown in the preceding section, these review mechanisms are embedded within defence institutions, operate as part of broader capability acquisition frameworks, and have evolved in parallel in response to challenges posed by emerging technologies, particularly AI-enabled systems. This shared institutional setting, the comparable functional and structural role of legal reviews, and a common appreciation of the challenges posed by cutting edge and emergent technologies – reflected in a concerted effort to address them – provide a plausible foundation for developing a joint review mechanism.

However, important deviations cast doubt on the viability and utility of such a mechanism. Divergence in the scope of legal reviews is only the first from a range of stumbling blocks to consider. While Pillar 2 identifies six technology-focused areas, not every technology falls in the scope of the legal review. As follows from the section above, all AUKUS partners conduct legal reviews of capabilities that, in their normal or intended use, are designed to inflict bodily harm or cause physical damage.⁸¹ As a practical matter,

therefore, where a Pillar II technology is integrated into a capability that is expected, in its ordinary employment, to result in damage, destruction, injury, or loss of life, that capability falls within the scope of legal review. For example, where an AI component is integrated into a weapon system to improve navigation or target identification and selection, that component falls within the scope of the legal assessment as part of the overall system.

Still, differences in the scope of review – most notably the US narrower approach to ‘methods of warfare’ and its exclusion of certain ‘means of warfare’ (such as platforms) – create uncertainty as to what capabilities are likely to be reviewed jointly. Most notably, such different scoping raises the risk that capabilities which the UK and Australia would ordinarily review within their legal review frameworks might, if developed jointly, fall outside the review ambit. Where the US addresses certain methods of warfare outside the formal legal review process, agreement on the appropriate alternative institutional locus and timing of such assessments would be required, adding complexity and creating potential gaps in oversight.

Notably, even within a single State, defining and classifying emerging military capabilities in a way that clearly signals to acquisition authorities which systems require legal review can be challenging. For example, the US Army reviews cyber weapons, a term that includes certain cyber capabilities,⁸² while the US Air Force prefers the broader term cyber capabilities, which may not necessarily be classified as weapons.⁸³ At the domestic level, this inconsistency in terminology could create confusion for an acquisition authority supplying the same cyber capability to both services. In a trilateral AUKUS context, similar terminological divergences – rooted in established national practices – could generate further ambiguity, complicating the agreement on what must be reviewed.

Further, divergences in the applicable legal standards, including the interpretation of SI/US rule and the treatment of environmental harm, may lead to different outcomes for the same capability. Last but not least, the fact that the three partners may be bound by different international and domestic law obligations and follow distinct interpretations of common standards presents a significant structural obstacle: any joint mechanism would struggle to produce outcomes that are simultaneously legally valid and operationally usable for all three States without either disregarding constraints applicable to some States or forcing a degree of alignment that States are neither legally required nor politically willing to accept.

A further structural constraint is the absence of a clearly defined **institutional platform to anchor a joint legal review mechanism**. The legal review mechanism would likely need to be embedded within a stable trilateral setting capable of consolidating expertise and ensuring continuity across projects. Restated, if Pillar 2 is indeed conceived as ‘one big factory across three nations,’ as Sharpe suggests, such a factory would necessarily require a dedicated office for a review authority within its premises.

At present, however, no such structure exists. AUKUS Pillar 2 remains characterised by fragmented governance: coordination occurs through evolving working groups, while capability development and funding decisions continue to be made largely through national processes.⁸⁴ This results in projects advancing on parallel, rather than truly integrated, tracks, with limited alignment of objectives, timelines, standards, and oversight practices. The problem is compounded by the lack of dedicated **funding** streams for Pillar 2 activities.⁸⁵ Rather than being supported by a distinct budget, Pillar 2 related work is absorbed into existing national programmes, reducing visibility, predictability, and strategic coherence. Today, these institutional and financial gaps make it difficult to envisage how a joint legal review mechanism could be effectively established, resourced, and sustained over time.

More fundamentally, the establishment of such a mechanism raises questions as to whether AUKUS States can discharge their legal review obligations through a mechanism not managed by their own authorities. While domestic review processes may, in practice, include exchange officers from partner States, ultimate responsibility remains with the reviewing State. In a joint mechanism, however, the allocation of responsibility would be less straightforward and would need to be expressly addressed. Particularly in case where the joint mechanism might be expected to take into account domestic prohibitions or restrictions applicable in one partner State that do not otherwise bind the others.⁸⁶ In such circumstances, participation in a joint review process would likely require enabling legislation – ie parliamentary authorisation – to address both the issue of responsibility and the incorporation of foreign domestic constraints. While this matter neither needs to be addressed in more detail nor resolved here, it underscores further legal and practical hurdles to the implementation of a joint review mechanism.

Lastly, even if a joint review mechanism were established, there remains a risk that one partner –most notably the US – could dominate its operation. A trilateral mechanism would need to function on the basis of equality, ensuring that each State retains an equal voice and that decision-making authority is balanced. Allowing one State to act as the de facto ultimate reviewer would risk reproducing asymmetries already evident in the export control domain, where the United States effectively set the parameters to which Australia and the United Kingdom adapted. Specifically, rather than developing a genuinely trilateral export control framework tailored to the objectives of AUKUS Pillar 2, the United States effectively insisted on the primacy of its domestic approach, long criticised by experts for its shortcomings.⁸⁷ Through the *National Defense Authorization Act for Fiscal Year 2024 (NDAA)*, the U.S. Congress created a national exemption from certain U.S. defence export licensing requirements for Australia and the United Kingdom, *conditional* upon the U.S. President certifying to Congress that both countries have implemented export control systems *comparable* to those of the United States.⁸⁸

Combined, these factors suggest that while a joint review mechanism may be conceptually appealing, its practical design and operation would be fraught with complexity and may ultimately require a more flexible, coordinated – rather than unified – approach. We discuss this approach in the next section.

Trilateral Legal Review Coordination Mechanism

While a joint legal review mechanism poses significant legal and practical challenges, relying on domestic review processes alone is equally unsatisfactory. As set out above, even though the AUKUS partners share a common core in the scope of legal reviews – covering weapons and their use – they diverge in the treatment of certain means and methods of warfare, and – in respect to the same military capability – may be bound by different legal obligations or adopt divergent interpretations of common norms.

A more feasible alternative may therefore lie in the establishment of a trilateral coordination mechanism (TCM). Such a mechanism – whether formal or informal – could serve as a structured forum for addressing issues arising in the context of national legal reviews. In particular, it could enable regular exchange of information and facilitate coordination on key aspects of the review processes and capability development. We suggest that the following key aspects be considered in the establishment of a TCM.

First, given that personnel responsible for conducting legal reviews in each AUKUS State rotate regularly and that review processes themselves are periodically revised, it is essential to establish **stable communication channels** among the partners. This would ensure a shared understanding of what each State reviews (scope), when reviews are conducted (timing), and how they are carried out (applicable legal standards). While much of this information is publicly available for Australia, the UK, and the US – and, in practice, the three have been engaging with one another on weapons reviews where necessary, making coordination comparatively straightforward – this may not be the case for some of the contemplated future AUKUS Pillar 2 participants. Should additional States join Pillar 2 – particularly those whose review mechanisms are less transparent or not well established, such as Japan or South Korea – ensuring mutual awareness of national legal review approaches and their differences becomes a critical starting point for effective coordination.

Second, **the timing of reviews** is another significant consideration. States may conduct legal reviews at different stages in the capability acquisition cycle and at varying speeds, with direct implications for jointly developed capabilities. For instance, where a State's acquisition framework links the release of funding to the achievement of specific milestones – some of which may depend on a positive legal review – the overall pace of a

joint programme may be determined by the slowest review process. Greater transparency among partners regarding such procedural requirements would facilitate better synchronisation of review timelines and more effective project planning.

Third is the issue of **information and data sharing**. The coordination of relevant information and data on the expected behaviour of the capability is critical for an effective legal review. Even within a single State, the conduct of meaningful legal reviews – particularly of complex, software-driven AI-enabled systems – depends on access to technical expertise and performance data that increasingly resides with industry rather than government.⁸⁹ In particular, legal review authorities would need to rely on structured collaboration with industry, especially for testing, evaluation, verification, and validation (TEVV) of capabilities. Yet such cooperation is frequently constrained by contractual limitations and proprietary concerns.⁹⁰

These challenges risk being amplified in a trilateral setting, where the number of actors – governmental, military, and industrial – increases, and where sensitivities relating to intellectual property and national security are heightened. In a joint development context, different components of a capability may be designed and built by different partners, such that relevant data is not automatically accessible to all participating States. As some Workshop participants observed, this issue may arise, for example, where one State subcontracts the development of a component to a private contractor that imposes strict non-disclosure obligations vis-à-vis third parties. In such a scenario, there is a risk that the other AUKUS partners may be unable to access the data necessary to conduct their own legal reviews.

We believe that data sharing constraints can, to a significant extent, be mitigated within a TCM. In particular, participating States can identify, at an early stage, the categories of information required for their respective legal reviews and communicate these requirements to their partners. This would allow the necessary administrative processes for access and transfer to be initiated early in the development, reducing the risk of downstream delays. In addition, where relevant information is held by private contractors in another State, obligations to generate and share specified data can be embedded contractually, ensuring that required information is available for the purposes of national legal reviews in other partner States.

Fourth, **coordination of (interim) review outcomes** is even more critical, since these outcomes directly impact the development of a capability. In principle, reviews may yield three types of outcomes: positive (the capability is cleared for use), negative (the capability cannot be used lawfully and is thus not cleared for use), or conditional (the capability is cleared for use in certain specific circumstances and not cleared for use in certain other circumstances). In practice, outright negative findings are rare, while conditional outcomes are more common.

In a context of joint capability development, these outcomes have direct consequences for partner participation. An interim negative review by one AUKUS partner would likely require that partner to withdraw from the project, as it could not lawfully support or contribute to the capability. A conditional outcome may limit a State's ability to support particular functions or uses of the system. Awareness of partners' (interim) review outcomes is therefore essential. Where a State can no longer support certain aspects of a capability, the joint capability development must either be adjusted to meet its legal requirements or continue without that partner's full participation. In the latter case, remaining partners would need to reassess the feasibility of the project, including funding arrangements and the redistribution of development and manufacturing responsibilities. A TCM would serve as a forum for the exchange of legal review outcomes, enabling their timely dissemination and facilitating coordinated responses among partner States. Importantly, this exchange need not involve the disclosure of full legal reviews. It may be limited to concise notifications identifying the aspect of the capability found to be unlawful or problematic and the resulting national position. Properly designed, such targeted information-sharing would avoid many of the sensitivities associated with any broader data disclosure.

Fifth is an issue of **security clearances**. The difficulty of aligning and recognising security clearance regimes across the three AUKUS partners may need to be addressed. Although there is broad structural similarity between national clearance and classification systems, there is no automatic mutual recognition, making the 'translation' of clearances a complex and often slow process.⁹¹ This has direct implications for legal review: access to classified technical data, operational concepts, and performance assessments – often essential for meaningful analysis – depends on personnel holding the appropriate clearances within each jurisdiction. While national systems, such as Australia's, are capable of allocating reviewers with the requisite clearance levels and managing sensitive or proprietary information through NDAs,⁹² extending this model across three States introduces additional layers of administrative, legal, and security coordination. The result is a potential bottleneck in information access, personnel deployment, and timely review.

At the same time, as suggested by one of the Workshop participants who conducted legal reviews for Australia in the past, security clearance requirements are unlikely to constitute an insurmountable barrier in practice. Rather, the challenge lies less in the existence of clearance regimes as such, and more in ensuring that coordination and mutual recognition processes within a TCM operate in a timely and sufficiently robust manner to enable rapid and meaningful access to the information required for legal review.

Seventh, a TCM may be formal or informal. An informal mechanism, not grounded in a specific legal arrangement, may be easier to establish and more adaptable to the evolving needs of a particular Pillar 2 capability development program. However, given that one of

the TCM's core functions would be the exchange of often classified information – such as performance data relevant to legal reviews or the outcomes of those reviews – which typically requires formal intergovernmental arrangements, the AUKUS partners may ultimately need to consider establishing the mechanism on a formal basis.

Lastly, TMC should not be confused with transparency measures often discussed and called for in the context of legal reviews. Specifically, data and outcome sharing have been proposed in public discussions of legal reviews as such transparency measures.⁹³ These measures are primarily directed outward, aiming to build broader confidence in the conduct and integrity of legal reviews.

The purpose of the TCM, however, is not to promote transparency beyond the participating States, but to minimise friction in the joint development of capabilities under AUKUS Pillar 2 by facilitating timely and targeted information exchange among partners. In this sense, TCMs are instruments of internal coordination rather than outward-facing transparency measures.

Conclusion

During the drafting of AP I, several States proposed the creation of an international mechanism to assess the lawfulness of new weapons. That proposal failed to attract sufficient support.⁹⁴ Instead, Article 36 was framed in a manner that left legal reviews a distinctly national undertaking. This reflected a pragmatic recognition: unless States retained discretion to design review processes compatible with their own institutional and organisational structures and defence demands and practices, formalised and systematic legal review regimes might never emerge at all. The framework adopted in AP I thus embodies a conscious compromise – imposing a shared legal obligation while preserving decentralised, nationally tailored implementation.

The creation of the AUKUS partnership, and in particular Pillar 2 cooperation on advanced capabilities, effectively reopens this question in a narrower (at this stage) trilateral setting. Yet, as this paper argues, even among three closely aligned partners, a fully joint multinational legal review mechanism is neither realistic nor desirable. Persistent divergences in how reviews are scoped, as well as in the interpretation and application of relevant legal standards, present structural obstacles that seem unlikely to be overcome. In this context, a TCM emerges, in our view, as the only credible and operationally viable pathway. To be sure, it is not without challenges, including potential constraints on technical data-sharing, the need to develop and align TEVV standards, to ensure smoother mutual recognition of security clearances, and to establish or recognise trusted third-party evaluators. These are not structural barriers, however, but practical issues that can be addressed. A TCM would preserve national review processes while enabling

targeted cooperation and structured information exchange. It would ultimately give effect to the spirit of Article 36 without demanding an unachievable degree of institutional alignment or convergence in the formulation and interpretation of legal standards.

Even if one accepts the view that, only a few years into the strategic partnership, both Pillars 1 and 2 are broadly where AUKUS leaders want them to be,⁹⁵ the time to think seriously about a trilateral legal review coordination mechanism is now. If such a mechanism is to support Pillar 2 effectively, the partners must already be considering how to design it and bring it into being.

About the Authors

Dr Natalia Jevglevskaja is an Honorary Lecturer at the Faculty of Law and Justice, UNSW Sydney. Her research focuses on the intersection of law, data, and technology in the financial, military, and security domains. She has published on legal reviews of emerging military technologies and on data and technology governance in financial services, including analysis of Australia's Consumer Data Right (CDR). Her work has appeared in leading academic and professional journals in Australia and internationally. She is the author of *International Law and Weapons Review* (Cambridge University Press, 2021) and co-author of *Customer Data Sharing Frameworks* (Routledge, 2024). In 2024–2025, Natalia served as a co-investigator on a multi-million-dollar DARPA contract awarded to UNSW Canberra, UNSW Sydney, and Arizona State University. The project contributed to DARPA's Autonomy Standards and Ideals with Military Operational Values (ASIMOV) program, which sought to develop objective, quantitative benchmarks for assessing the ethical complexity of future autonomy use cases and the readiness of autonomous systems to operate in those contexts in accordance with military operational values. Natalia is also a Director of the Asia-Pacific Institute for Law and Security and an Associate Fellow of the Higher Education Academy, UK.

Renato Wolf is an Honorary Research Fellow at the University of Queensland Law School. He served as a legal adviser to the Swiss Armed Forces for nearly ten years and was deployed to Kosovo and Bosnia-Herzegovina and to the United Nations Headquarters. He holds a Bachelor of Law and a Master of Law from the University of Bern, Switzerland, as well as a Master of Arts in War in the Modern World from the King's College London. He completed a PhD on the legal review of autonomous weapons systems at the University of Queensland.

ENDNOTES

- 1 AUKUS is a trilateral security partnership announced in September 2021 between Australia, the United Kingdom and the United States structured around two principle lines of effort: Pillar I concerns Australia's acquisition of conventionally armed, nuclear-powered submarines; Pillar II concerns cooperation on advanced defence capabilities, including cyber, AI, quantum, undersea systems, hypersonics and electronic warfare. See Luke A. Nicastro, "AUKUS Pillar 2 (Advanced Capabilities): Background and Issues for Congress" (Congressional Research Service, May 2024) 4 <<https://sgp.fas.org/crs/row/R47599.pdf>>.
- 2 The White House, "Joint Leaders Statement on AUKUS," September 15, 2021, <https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2021/09/15/joint-leaders-statement-on-aucus/>.
- 3 William Greenwalt and Tom Corben, "AUKUS Enablers? Assessing Defence Trade Control Reforms in Australia and the United States" (Sydney: United States Studies Centre at the University of Sydney, August 2024), 8, <https://cdn.sanity.io/files/ooh1fq7e/production/152c48a93f6ac0a7bc04719e0cff612a-590c8ac5.pdf/Assessing-defence-trade-control-reforms-in-Australia-and-the-United-States.pdf>.
- 4 Amy McDonnell, "Uncertified: The Implications of the US Department of State's Decision on Australia's ITAR Exemption" Export Controls Australia Group, 1 May, 2024, <https://www.ecag.com.au/post/uncertified-the-implications-of-the-us-department-of-state-s-decision-on-australia-s-itar-exemption>.
- 5 Nicastro, "AUKUS Pillar 2", 4.
- 6 McFarland, T., Assaad, Z., "Legal reviews of *in situ* learning in autonomous weapons". *Ethics Inf Technol* 25, 11 (2023). Sanders, L., & Copeland, D. (2020, November 27). Developing an Approach to the Legal Review of Autonomous Weapon Systems. ILA Reporter. <https://ilareporter.org.au/2020/11/developing-an-approach-to-the-legal-review-of-autonomous-weapon-systems> lauren-sanders-and-damian-copeland/; Klaudia Klonowska, "Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare," Asser Research Paper No. 2021-02 (The Hague: T.M.C. Asser Institute for International & European Law, March 17, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3823881.
- 7 Austin Wyatt et al, "Towards AUKUS Collaboration on Responsible Military Artificial Intelligence", RAND Australia, 2024 18, www.rand.org/content/dam/rand/pubs/research_reports/RRA3000/RRA3079-1/RAND_RRA3079-1.pdf.
- 8 See section C below.
- 9 James Palmer, "AUKUS needs a third pillar: space", Australian Strategic Policy Institute, 27 June, 2025, <https://www.aspistrategist.org.au/aukus-needs-a-third-pillar-space/>.
- 10 Nicastro, "AUKUS Pillar 2", 1.
- 11 The White House, "Joint Leaders Statement".
- 12 United Kingdom Government, "Fact Sheet: Implementation of the Australia – United Kingdom – United States Partnership (AUKUS)", 5 April, 2022, <https://www.gov.uk/government/publications/implementation-of-the-australia-united-kingdom-united-states-partnership-aukus-fact-sheet/1a1ad299-e2c8-4ef5-82e3-9af0e057fe9a>.
- 13 Nicastro, "AUKUS Pillar 2", 1.
- 14 Michael Shoebridge, "What is AUKUS and what is it not?", 2021 ASPI Strategist, 4, https://ad-aspi.s3.ap-southeast-2.amazonaws.com/2021-12/What%20is%20AUKUS%20and%20what%20is%20it%20not.pdf?VersionId=uAcnyQeum5lvEdiJYxSRW5SubMT2AMIV&utm_source=chatgpt.com.
- 15 Australian Strategic Policy Institute, "AUKUS Relevant Technologies: Top 10 Country Snapshot", 2023 1 <https://ad-aspi.s3.ap-southeast-2.amazonaws.com/2023-06/PB69-CriticalTechTracker-AUKUS%20relevant%20technologies%20top%2010%20country%20snapshot.pdf>.
- 16 Ibid.
- 17 Central Committee of the Communist Party of China (CPC) and People's Republic of China State Council, "Outline of the National Innovation-Driven Development Strategy Issued by the CPC Central Committee and the State Council," trans. Center for Security and Emerging Technology, 11 December, 2019, <https://cset.georgetown.edu/publication/outline-of-the-national-innovation-driven-development-strategy/>.

- 18 See also Arthur R. Kroeber, “Unleashing “New Quality Productive Forces”: China’s Strategy for Technology-led Growth”, The Brookings Institution, 4 June, 2024, <https://www.brookings.edu/articles/unleashing-new-quality-productive-forces-chinas-strategy-for-technology-led-growth/>.
- 19 Sarah Tzinieris et al, “Shedding Light on Chinese Thinking on AUKUS”, Pacific Focus Vol. XXXIX, No. 3 (2024): 504, <https://doi.org/10.1111/pafo.12262>.
- 20 Kim Beazley, “We need a third pillar of AUKUS: critical minerals”, Australian Strategic Policy Institute, 23 May, 2024, <https://www.aspistrategist.org.au/we-need-a-third-pillar-of-aukus-critical-minerals/>.
- 21 Adam Thorn, “To infinity and beyond! Former PM calls for AUKUS to expand to space”, Defence Connect, 28 May 2025, <https://www.defenceconnect.com.au/joint-capabilities/16125-to-infinity-and-beyond-former-pm-calls-for-aukus-to-expand-to-space>; See also Palmer, “AUKUS needs a third pillar”.
- 22 See Natalia Jevglevskaja, “AUKUS Pillar 2: Legal Implications of Sharing Advanced Technology” [pp xy], in *Armed Conflict and International Law in the Indo-Pacific Region*, Douglas Guilfoyle and Jennifer Maddocks, Oxford University Press, 2026, doi: 10.1093/oso/9780197844540.003.0014. See also Lisa Curtis and Ryan Claffey, “Congressional support key for sustaining momentum of US, UK, Australia partnership”, The Hill, 02 February, 2026, <https://thehill.com/opinion/national-security/5717755-aukus-defense-industrial-base/>.
- 23 Tom Corben and Sophie Mayo, “Patience, please. AUKUS Pillar 2 is indeed making progress”, Australian Strategic Policy Institute, 19 April, 2024, <https://www.aspistrategist.org.au/patience-please-aukus-pillar-2-is-indeed-making-progress/>; Peter Tesch, “AUKUS: Two pillars, three fallacies”, The Lowy Institute, 13 March, 2025, <https://www.loyyinstitute.org/the-interpreter/aukus-two-pillars-three-fallacies>; Molly Donohue Magee, “AUKUS Tech Development A Marathon, Not Sprint”, National Defense Magazine, 18 July, 2024, <https://www.nationaldefensemagazine.org/articles/2024/7/18/aukus-tech-development-a-marathon-not-sprint>.
- 24 American Chamber of Commerce in Australia and Business Council of Australia partnership. “Australia’s AUKUS Pillar II Opportunity”, November, 2025, 5, <https://www.bca.com.au/wp-content/uploads/2025/11/Australias-AUKUS-Pillar-II-Opportunity-BCA-AmCham-report-WEB.pdf>.
- 25 Peter Dean and Alice Nason, “AUKUS Pillar II Is Failing in Its Mission”, War on the Rocks, 2 June, 2025, <https://warontherocks.com/2025/06/aukus-pillar-ii-is-failing-in-its-mission-it-needs-its-own-optimal-path-way/#:~:text=Pillar%20I%20submarine%20cooperation%20remains,Class%20Justin%20Yarborough%20via%20DVIDS>; George Allison, “AUKUS Pillar Two ‘lacks focus and delivery’ say MPs”, UK Defence Journal, 14 December 2025, <https://ukdefencejournal.org.uk/aukus-pillar-two-lacks-focus-and-delivery-say-mps/>.
- 26 Nicola Macaulay, “Is NZ defence and intelligence policy aligning with AUKUS in all but name?”, The Conversation, February 4, 2026, <https://theconversation.com/is-nz-defence-and-intelligence-policy-aligning-with-aukus-in-all-but-name-274609>; Adam Broinowski, “AUKUS Pillar 2, Japan and South Korea”, Parliamentary Library, August 15, 2024, https://parlinfo.aph.gov.au/parlInfo/download/library/prspub/9893294/upload_binary/9893294.pdf; Alexander Gray and Doug Stokes, “From AUKUS to CAUKUS: The Case for Canadian Integration”, Legatum Institute, February 2024, <https://www.prosperity.com/wp-content/uploads/2024/02/AUKUS-PILLAR-2.pdf>.
- 27 See also Simon McKenzie and Eve Massingham, “AUKUS: The Regulation of the Ocean and the Legal Dangers of Working Together” *Ocean Yearbook Online* 37, No 1 (2023) 148-152.
- 28 *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).
- 29 The reference to assessment “prior” to development or acquisition should not be understood as implying that the review obligation is exhausted once a capability is fielded. Rather, it identifies the minimum point at which legal scrutiny must occur. Where circumstances so require – such as where a system is materially modified, thereby producing effects different from those originally reviewed, or is deployed in novel ways – further review may be necessary after deployment. See Netta Goussac et al., “*Enhancing the Legal Review of Autonomous Weapon Systems: Report of an Expert Meeting (Sydney, 28–30 March 2023)*”, Law and the Future of War Research Group, TC Beirne School of Law, The University of Queensland, 1 May, 2023, 11, <https://doi.org/10.14264/2bbfd31>
- 30 Natalia Jevglevskaja, *International Law and Weapons Review: Emerging Military Technology under the Law of Armed Conflict* (Cambridge University Press, 2021) ch 5.5.

- 31 See Goussac et al., “*Enhancing the Legal Review*”, sec 2.6; ‘The review examined the Winchester pump action shotgun firing buckshot pellets to be employed in compliance with the customary law rule prohibiting the use of particularly injurious weapons (now codified in Art. 35(2) AP I)’, see Renato Wolf et al., “*Advancing the Legal Review of Autonomous Weapon Systems: Report of an Expert Meeting (Sydney, 16–18 April 2024)*”, Law and the Future of War Research Group, TC Beirne School of Law, The University of Queensland, 1 September, 2024, 12, <https://doi.org/10.14264/c41c6fb>.
- 32 See Wolf et al. “*Advancing the Legal Review*”, 12.
- 33 Australian Department of Defence, “Defence Instruction (General) OPS 44-1: Legal Review of New Weapons”, 2 June, 2005.
- 34 Australian Department of Defence, “Defence Article 36 Reviews of New Weapons Guide”, May, 2020.
- 35 Department of Defence, “*Australia’s Guide to the Legal Review of New Weapons, Means or Methods of Warfare*”, 10 April, 2024, https://legalreviewportal.org/files/au/2024-04-10-adf_guide_to_art_36_legal_review.pdf.
- 36 Department of Defence, “AF172: Article 36 Legal Review Report”, 16 November, 2023.
- 37 Wolf et al. “*Advancing the Legal Review*”, 11.
- 38 Damian Copeland and Rain Liivoja, “*Progressing the Legal Review of Autonomous Weapon Systems: Report of an Expert Meeting (Geneva, 10-11 March 2025)*”, The Asia-Pacific Institute for Law and Security, September, 2025, 9, <https://apils.org/2025/09/04/report-of-the-2025-expert-meeting-on-the-legal-review-of-aws/>.
- 39 Copeland and Liivoja, “Progressing the Legal Review” 9. See also United Kingdom Ministry of Defence Development, Concepts and Doctrine Centre, “*UK Weapon Reviews*” (2016), https://assets.publishing.service.gov.uk/media/5a80bf5f40f0b62305b8cec5/20160308-UK_weapon_reviews.pdf.
- 40 Copeland and Liivoja, “Progressing the Legal Review”, 9.
- 41 Wolf et al. “*Advancing the Legal Review*”, 11.
- 42 Copeland and Liivoja, “Progressing the Legal Review”, 10.
- 43 Wolf et al. “*Advancing the Legal Review*”, 12.
- 44 Copeland and Liivoja, “Progressing the Legal Review”, 10.
- 45 Copeland and Liivoja, “Progressing the Legal Review”, 10.
- 46 A collaborative organisation that brings together industry, academia, government and defence stakeholders across the three AUKUS countries to promote cooperation, innovation and economic opportunities arising from the partnership, see “AUKUS Forum”, accessed 20 March, 2026, <https://aukusforum.com/>.
- 47 Allyson Park, “U.S., U.K., Australia Still Ironing Out Tech, Info Sharing Kinks”, National Defense Magazine, 7 March, 2025, <https://www.nationaldefensemagazine.org/articles/2025/3/7/us-uk-australia-still-ironing-out-tech-info-sharing-kinks>.
- 48 Note that in the US, certain technology, such as AWS, is subject to policy-level assessment at the Department of Defense level in addition to a legal review of the weapon system conducted by the respective service. See Department of Defense. “DoD Directive 5000.01, The Defense Acquisition System” United States of America, July 28, 2022 <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/500001p.pdf>; Department of Defense. “Directive Number 3000.09: Autonomy in Weapon Systems.” United States of America, May 8, 2017. <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>; Wolf et al. “*Advancing the Legal Review*” 12-13).
- 49 Vincent Boulanin and Maaïke Verbruggen, “*Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies*,” Stockholm International Peace Research Institute, December, 2017, 15, https://www.sipri.org/sites/default/files/2017-12/sipri_bp_1712_article_36_compendium_2017.pdf.
- 50 See also Netta Goussac and Rain Liivoja, “Legal Reviews of Military Artificial Intelligence Capabilities”, Asia-Pacific Institute for Law and Security, 25 August, 2025, <https://apils.org/2025/08/25/legal-reviews-of-military-artificial-intelligence-capabilities/>.
- 51 However, see also our comment in FN 28 above.
- 52 Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems “*Australia’s System of Control and Applications for Autonomous Weapon Systems*”, 26 March, 2019, <https://undocs.org/CCW/GGE.1/2019/WP.2/Rev.1>.
- 53 Ibid, para 3.

- 54 Goussac et al., “Enhancing the Legal Review”, 11.
- 55 DoD Directive 2311.01: DoD Law of War Program (2 July 2020) s 1.4. See also Andrew McKee and Christina Colclough, “Legal Reviews of Autonomous Weapons at the Tactical Edge,” *Articles of War*, Lieber Institute West Point, November 12, 2025, <https://lieber.westpoint.edu/legal-reviews-autonomous-weapons-tactical-edge/>.
- 56 This understanding draws on in-person discussions with officials responsible for legal reviews in Australia, the UK and the US, held during the Sydney Expert Meetings on Autonomous Weapon Systems. See generally Goussac et al., “*Enhancing the Legal Review*” and Wolf et al. “*Advancing the Legal Review*”.
- 57 Department of Defence, “Australia’s Guide to the Legal Review”, sec. 5(a)–(b).
- 58 United Kingdom Ministry of Defence Development, Concepts and Doctrine Centre, “UK Weapon Reviews”, 4-5. (While the current Joint Services Publication is classified, there are no indications in the publicly available material that the UK has changed its treatment of ‘weapons / means’).
- 59 United States, “*Department of the Air Force Instruction 51-401: The Law of War*,” 3 August, 2018. pt 2, para. 5 and Glossary, https://static.e-publishing.af.mil/production/1/af_ja/publication/afi51-401/afi51-401.pdf; United States, “*Department of the Army Regulation 27-53: Legal Review of Weapons and Weapon Systems*”, 23 September, 2019, Glossary; United States, “*Secretary of the Navy Instruction 5000.2F: Defense Acquisition System and Joint Capabilities Integration and Development System Implementation*”, 26 March, 2019, encl 3, s 10(e)(1).
- 60 United States, “*Department of the Air Force Instruction 51-401*”, para. 5 and Glossary (emphasis added).
- 61 United States, “*Department of the Air Force Instruction 51-401*”, pt 2, para. 5 and Glossary; United States, “*Department of the Army Regulation 27-53*”, sec 5(d)(2); United States, “*Secretary of the Navy Instruction 5000.2F*”, encl 3, s 10(e)(3).
- 62 Department of Defence, “Australia’s Guide to the Legal Review”, Part I, s 5(c).
- 63 Wolf et al. “*Advancing the Legal Review*”, 11.
- 64 United States, “*Department of the Air Force Instruction 51-401*”, pt 2, sec 6.1.2; United States, “*Department of the Army Regulation 27-53*”, sec 8(b); United States, “*Secretary of the Navy Instruction 5000.2F*”, encl 3, s 10(f)(2).
- 65 This understanding is informed by in-person discussions with officials responsible for legal reviews in the United States, held during the Sydney Expert Meetings on Autonomous Weapon Systems. See generally Goussac et al., “*Enhancing the Legal Review*” and Wolf et al. “*Advancing the Legal Review*”.
- An example is Department of Defense Directive 3000.09 on *Autonomy in Weapon Systems*, which requires that such systems ‘be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force’, and that the system design incorporates the capabilities necessary to support this requirement before development proceeds. See Department of Defense, ‘Directive Number 3000.09: Autonomy in Weapon Systems’ <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>, Art. 4(1)(c)(1).
- 66 Art. 35(2) AP I and International Committee of the Red Cross Customary IHL Database, “Rule 70. Weapons of a Nature to Cause Superfluous Injury or Unnecessary Suffering,” accessed May 15, 2026.
- 67 *Legality of the Threat or Use of Nuclear Weapons (Advisory Opinion)* [1996] ICJ Reports 226, p. 35, para. 78.
- 68 Rain Liivoja, “Protecting Warfighters from Superfluous Injury and Unnecessary Suffering”, in “*Civility, Barbarism and the Evolution of International Humanitarian Law*”, ed. Matt Killingsworth and Tim McCormack (University of Oxford Press 2024), 189.
- 69 US Department of the Army (1956). Field Manual (FM) 27-10: The Law of Land Warfare, Washington, DC: Department of the Army.
- 70 W. Hays Parks, “Conventional Weapons and Weapons Reviews”. 8 *Yearbook of International Humanitarian Law* 55, 86-87 footnote 123 (2005); Richard R Baxter, “Conventional Weapons Under Legal Prohibitions”, 1 *International Security* 42, 43 (1977).
- 71 U.S. Department of Defense, Department of Defense Law of War Manual (Washington, DC: Office of General Counsel, 2015; updated July 2023), section 6.6.

- 72 US Law of War Manual, section 6.6.1. For an explanation as to why the US is in the minority on this, see *ibid*, Rain Liivoja, “Protecting Warfighters from Superfluous Injury and Unnecessary Suffering”, 188-189.
- 73 Cf Liivoja, “Protecting Warfighters from Superfluous Injury and Unnecessary Suffering”, 188.
- 74 Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law—Volume I: Rules*, (Cambridge University Press 2005), Rule 45; See also Saeed Bagheri and Gerhard Kemp, “Damage to and Destruction of the Natural Environment: Terraforming Warfare in Gaza and Accountability for Ecocentric Crimes”, *Journal of International Criminal Justice* 23, no 3-4 (2025): 436, <https://doi.org/10.1093/jicj/mqaf049>.
- 75 United States, *Department of Defense Law of War Manual*, (Office of General Counsel, July 2023), sec. 6.10.3.1.
- 76 United States, “Department of the Air Force Instruction 51-401”, pt 2, para. 5; United States, “Department of the Army Regulation 27-53”, sec. 6(a); United States, ‘Secretary of the Navy Instruction 5000.2F”, encl. 3, s 10(a).
- 77 Department of Defence, “Australia’s Guide to the Legal Review”, part 1 section 2 and part 2 sections 10, 19-21 and 24.
- 78 Copeland and Liivoja, “Progressing the Legal Review”, 9-10.
- 79 This follows from the foundational principle of international law – the sovereign equality of States, Article 2(1) of the UN Charter, Charter of the United Nations, June 26, 1945, 1 U.N.T.S. XVI. See also *Treatment of Polish Nationals and Other Persons of Polish Origin or Speech in the Danzig Territory, Advisory Opinion*, 1932, P.C.I.J., Series A/B, No. 44, pp. 24-25. See also *Lotus*, Judgment No. 9, 1927, P.C.I.J., Series A, No. 10, p. 24. See also International Law Commission, *Draft articles on Responsibility of States for Internationally Wrongful Acts, with commentaries*, UN Doc A/56/10, Article 3, para. 2 https://legal.un.org/ilc/texts/instruments/english/commentaries/9_6_2001.pdf.
- 80 *Convention on Cluster Munitions*, opened for signature 3 December 2008, 2688 UNTS 39 (entered into force 1 August 2010).
- 81 Natalia Jevglevskaia, *International Law and Weapons Review: Emerging Military Technology under the Law of Armed Conflict*, (Cambridge University Press 2021), ch. 4.1.
- 82 United States, “Department of the Army Regulation 27-53”, glossary and sec. 8.
- 83 United States, “Department of the Air Force Instruction 51-401”, Attachment 1 and part 2 section 7.
- 84 Peter Dean and Alice Nason, “AUKUS Pillar II Is Failing in Its Mission”, War on the Rocks, 2 June, 2025, <https://warontherocks.com/2025/06/aukus-pillar-ii-is-failing-in-its-mission-it-needs-its-own-optimal-path-way/#:~:text=PIllar%20I%20submarine%20cooperation%20remains,Class%20Justin%20Yarborough%20via%20DVIDS>.
- 85 *Ibid*.
- 86 See section C(2)(ii) above discussing the standard of review.
- 87 The criticisms focus in particular on the substance and operation of the *International Traffic in Arms Regulations* (ITAR) administered by the Directorate of Defence Trade Controls (DDTC) [ITAR, 22 CFR §§120-130]. The prominent ‘eight deadly sins’ if ITAR include ‘universality and non-materiality’, ‘extraterritoriality’, ‘non-discrimination between different classes of recipients’, ‘transactional process-compliance’, ‘knowledge taint’, ‘non-reciprocity’, and ‘unwarranted predictability’. On their detailed substantiation, see William Greenwalt and Tom Corben, “*Breaking the Barriers: Reforming US Export Controls to Realise the Potential of AUKUS*”, *United States Studies Center*, 2023, 10-16, <https://www.ussc.edu.au/breaking-the-barriers-reforming-us-export-controls-to-realise-the-potential-of-aukus>.
- 88 S 1343 NDAA (amending Section 38 of the Arms Export Control Act (22 U.S.C. 2778)). This certification was issued to Congress on 15 August 2024, see The US Department of Defence, “AUKUS Defense Trade Integration Determination”, 15 August, 2024, <https://2021-2025.state.gov/aukus-defense-trade-integration-determination>.
- 89 Goussac and Liivoja, “Legal Reviews of Military Artificial Intelligence Capabilities”.
- 90 Daniel Fiott, “Military-technological cooperation across the Euro-Atlantic and Indo-Pacific”, War on the Rocks, 19 January, 2024, <https://warontherocks.com/2024/01/military-technological-cooperation-across-the-euro-atlantic-and-indo-pacific/>.

- 91 American Chamber of Commerce in Australia and Business Council of Australia partnership. “Australia’s AUKUS Pillar II Opportunity”, November, 2025, 5, <https://www.bca.com.au/wp-content/uploads/2025/11/Australias-AUKUS-Pillar-II-Opportunity-BCA-AmCham-report-WEB.pdf>.
- 92 See Department of Defence, “Australia’s Guide to the Legal Review”, Part I, sections 21-22.
- 93 Wolf et al., “Advancing the legal review”, p. 19.
- 94 See Natalia Jevglevskaja, *International Law and Weapons Review: Emerging Military Technology under the Law of Armed Conflict* (Cambridge University Press, 2021) ch 5.1. See also Netta Goussac and Rain Liivoja, “Legal Reviews of Military Artificial Intelligence Capabilities”, Asia-Pacific Institute for Law and Security, 25 August, 2025, <https://apils.org/2025/08/25/legal-reviews-of-military-artificial-intelligence-capabilities>.
- 95 Kris Osborn, “AUKUS Nations Shift Priorities From R&D to Deliverable Technologies,” National Defense Magazine, June 6, 2025, <https://www.nationaldefensemagazine.org/articles/2025/6/6/aukus-nations-shift-priorities-from-rd-to-deliverable-technologies>.

/ PREDICTABILITY IN THE GOVERNANCE OF AUTONOMOUS WEAPON SYSTEMS¹

Rain Liivoja

Introduction

‘Predictability’ features as a key concept in discussions about the regulation of autonomous weapon systems (AWS). Most prominently, the International Committee of the Red Cross (ICRC) has been calling on states to address the risks of AWS by adopting a new legally binding instrument that expressly rules out unpredictable AWS.² Many states have supported this approach.³ Some have put forward alternative proposals that, instead of including unpredictability as an element of a prohibition, have identified predictability as a positive requirement for an acceptable AWS.⁴ Multiple other states, however, have not been persuaded.

Consequently, the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE) has been struggling to find agreement on language addressing this issue in the ‘set of elements of an instrument’, which the group has been drafting under its 2024–2026 mandate.⁵ Early iterations of this document would have required states to ensure that the effects of AWS are ‘adequately predictable, reliable, traceable and explainable’.⁶ Having been opposed by several states,⁷ this language was dropped from subsequent versions.⁸ It then returned as a proposed requirement to ensure that:

[AWS] ... in their identification, selection, and engagement functions, are adequately predictable and reliable in operation, and that their functioning and effects are traceable and explainable to those responsible for their development and use.⁹

The rejection of the language referring to predictability by several states does not necessarily mean that they do not see predictability as a desirable attribute of AWS. Rather, they appear to have questions about the precise meaning of this and associated

terms. Some states therefore suggest that these terms require further discussion and clarification before being considered for elevation into binding standards.¹⁰ While some of these calls for further terminological debate probably aim to draw out the debate, they are not entirely unfounded. In particular, while multiple national policy documents concerning the responsible use of military AI rely on notions such as reliability, traceability and explainability, they do not use or explain the concept of predictability.¹¹ Australia's 'policy settings' on military AI are a case in point.¹² Predictability is not one of their 'values-based principles' and, indeed, the concept makes no appearance in the document.

This paper considers predictability as a possible normative standard for inclusion in a regulatory instrument on AWS and, by extension, in the governance of military AI. The paper proceeds as follows. After a few notes about vocabulary, it outlines the so-called 'predictability problem' with autonomous and AI-enabled systems. Then, it turns to the way predictability has been defined and conceptualised in the literature. Next, it traces the way in which references to predictability have evolved in the GGE (including in relation to existing law) and the emergence of predictability as a possible regulatory standard. Finally, it comments on the advantages and disadvantages of identifying predictability as a discrete standard. The paper concludes that, rather than including predictability as an undefined concept in any new regulatory instrument, it would be preferable to incorporate into such an instrument an articulation of what predictability substantively requires.

Note on Terminology

Before turning to predictability specifically, a brief clarification about the notion of AWS may be helpful, especially as the 'characterisation' of AWS has been one of the sticking points throughout the work of the GGE.

This paper adopts a broad understanding wherein AWS are weapon systems that, after activation, can select and engage targets without intervention by a human operator. This broadly aligns with definitions used by the United States and the ICRC.¹³

The characterisation contemplated by the GGE has built upon the same understanding but added additional elements owing to the specific trajectory of debate within the group and given its mandate to deal with 'lethal' AWS (often branded with the somewhat unfortunate acronym 'LAWS'). First, the GGE has introduced target identification as a discrete critical function of an AWS and added a clarification as to its meaning.¹⁴ This paper implicitly follows the ICRC's approach whereby the system's ability to select targets presumes and incorporates the capacity to search for, detect, identify and track targets.¹⁵ Second, the GGE's characterisation also unpacks the notion of 'weapon system', which has caused difficulties for some states.¹⁶ Third, the GGE's characterisation contains a

savings clause to ensure that AWS that ‘can be used in a way that does not result in loss of life’ do not fall outside the scope of the instrument.¹⁷ The latter two issues need not be addressed here in detail.

The understanding of AWS adopted in this paper is relatively technology-neutral: it specifies neither the technological means for achieving autonomous functionality nor the precise degree of autonomy in the target selection and engagement functions. On this reading, a landmine is a primitive AWS, and a close-in weapon system (such as Phalanx) is a somewhat more advanced AWS. The GGE has, however, expressly reserved the possibility of excluding certain types of systems from its characterisation, and thus from the scope of its instrument.¹⁸ If this possibility were to be taken up, it would be most likely to affect weapon systems with limited autonomy in their critical functions—for example, automatic contact mines¹⁹ and precision-guided munitions.

In common parlance, AI refers to digital technologies and tools capable of performing tasks normally requiring human intelligence.²⁰ More formally, AI has been defined as the discipline focused on the research and development of mechanisms and applications of ‘AI systems’, which in turn are engineered systems that generate outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives.²¹ Currently, machine learning is the most common subset of AI. This approach is focused on algorithms that can ‘learn’ the patterns of training data and, subsequently, make inferences about new data.²² As a general-purpose technology, AI has a multitude of uses in the military domain across strategic, operational and tactical levels.²³ Among other things, AI can be a powerful ‘enabler’ of autonomy in weapon systems.²⁴ AI creates ‘improvement opportunities in all application areas of autonomy in weapon systems, from target recognition to navigation’²⁵ and allows autonomous targeting to ‘leave its former military niche applications and become adoptable across the board’.²⁶

Therefore, an AWS may be AI enabled, but it might not be. Conversely, AI may be used to enable some functionality in AWS, but it might not be. That having been said, many of the concerns expressed by states and civil society actors with respect to AWS relate more specifically to AI-enabled AWS, as demonstrated by the pervasiveness of AI-specific concepts in the LAWS GGE discussions.²⁷ Thus, much of the discussion around the regulation of AWS would also have some bearing on the governance of military AI generally.

The Predictability Problem

Writing in 2004, Andreas Matthias observed that, traditionally, the manufacturer or operator of a machine is held morally and legally responsible for the consequences of its operation.²⁸ He then identified an ‘accountability gap’:

Autonomous, learning machines, based on neural networks, genetic algorithms and agent architectures, create a new situation, where the manufacturer/operator of the machine is in principle not capable of predicting the future machine behaviour any more, and thus cannot be held morally responsible or liable for it.²⁹

A few years later, Rob Sparrow made a similar point regarding AWS:

While [artificially intelligent weapon systems] will be programmed to make decisions according to certain rules, in important circumstances their actions will not be predictable. ... The more the system is autonomous then the more it has the capacity to make choices other than those predicted or encouraged by its programmers. At some point then, it will no longer be possible to hold the programmers/designers responsible for outcomes that they could neither control nor predict.³⁰

While addressing the ‘accountability gap’, these two papers also offered an articulation of what has now become known as the ‘predictability problem’ (or the ‘unpredictability problem’). At its core, this is a simple but important difficulty: when a system operates without direct real-time human intervention—by sensing its environment, analysing the data, and computing and taking the appropriate course of action—it becomes harder for humans to foresee exactly how it will behave in all circumstances.

Some contemporary authors trace the articulation of the predictability problem back to Norbert Wiener, the founder of cybernetics.³¹ The specific point of reference is a 1959 public lecture in which Wiener argued that ‘machines can and do transcend some of the limitations of their designers, and that in doing so they may be both effective and dangerous’.³² In that lecture, Wiener linked unpredictability to the ability of a system to learn. Specifically, he suggested that ‘[a]s machines learn they may develop unforeseen strategies at rates that baffle their programmers’.³³ For Wiener, it seems to have been the speed and frequency of unforeseen strategies that potentially perplexed humans and became the source of concern, rather than the unforeseeability of individual behaviours more broadly.³⁴

Arthur Samuel—the man who coined the term ‘machine learning’—published a refutation of Wiener’s central claims. He posited that ‘[t]he machine is not a threat to mankind, as some people think’, and went on to argue ‘that machines cannot possess originality in the sense implied by Wiener and that they cannot transcend man’s intelligence’.³⁵

Interestingly, Samuel conceded that neural networks might constitute an exception to this conclusion. He noted that '[s]ince the internal connections [of neural networks] would be unknown, the precise behavior of the nets would be unpredictable and, therefore, potentially dangerous'.³⁶

Despite the rather longstanding recognition of what appears to be a pivotal problem with AI,³⁷ in-depth examinations have been rather scarce until recently. Roman Yampolskiy, a pioneer of AI safety research, lamented in 2020 about:

[the] poorly understood concept of unpredictability of intelligent systems, which limits our ability to understand the impact of intelligent systems we are developing and is a challenge for software verification and intelligent system control, as well as AI Safety in general.³⁸

He was not alone in identifying a lack of attention to the concept. At roughly the same time, Arthur Holland Michel noted:

[F]or the most part in the discourse on [AWS] and military AI, predictability and understandability have not yet been treated with the kind of detailed foregrounding that befits an issue of such importance and complexity.³⁹

The situation has, meanwhile, improved. For example, Yampolskiy himself went on to define the unpredictability of AI as 'our inability to precisely and consistently predict what specific actions an intelligent system will take to achieve its objectives, even if we know terminal goals of the system'.⁴⁰ To give another example, Tobias Vestner and Altea Rossi discussed unpredictability in terms of 'inherent uncertainty as to how systems will behave once deployed and respond to changing and complex environments'.⁴¹

Meaning of Predictability

Predictability is, of course, the inverse of unpredictability. It is a common enough word such that, when used in an international law instrument, it should presumptively be read based on its 'ordinary meaning'.⁴² This can be derived from dictionaries. The *Oxford English Dictionary* defines predictability as '[t]he fact or condition of being predictable; the extent to which something can be predicted or used for prediction'.⁴³ The adjective 'predictable' means '[a]ble to be predicted or foretold',⁴⁴ and the verb 'predict' means '[t]o state or estimate, esp[ecially] on the basis of knowledge or reasoning, that (an action, event, etc.) will happen in the future or will be a consequence of something'.⁴⁵

ICRC's foundational work on the predictability of AWS drew expressly on a similar dictionary definition.⁴⁶ On that basis, the ICRC conceptualised the predictability of AWS as 'knowledge of how the weapon system will function in any given circumstances of

use, including the effects that will result'.⁴⁷ The reference to 'knowledge' creates some challenges. For one, there is a live philosophical debate about whether one can have 'knowledge' about what will happen in the future. In more technical terms, this is:

the question whether future contingents are knowable, that is, whether one can know that things will go a certain way even though it is possible that things will not go that way.⁴⁸

Criminal law—where the knowledge of a specific consequence of an act is often an element of an offence—offers something of a way out. Thus, knowledge of a consequence may mean awareness that this consequence will occur 'in the ordinary course of events'.⁴⁹ But this becomes challenging to apply to the functioning of a weapon system generally or with respect to 'any given circumstances of use'. In any event, the general knowledge of the functioning of a system and the ability to predict its effects under specific conditions seem to be distinct issues.

Mariarosaria Taddeo circumvents some of these problems by simplifying the question: for her, predictability 'indicates the degree to which one can answer the question: what will an AI system do?'.⁵⁰ This understanding is elegant in its simplicity but perhaps slightly vague. Holland Michel characterises predictability as 'the extent to which a system's outputs or effects can be anticipated'.⁵¹ This comes reasonably close to the dictionary definition mentioned earlier but, instead of focusing on the ability to 'estimate' what will happen in the future, it uses the concept 'anticipate', which is more common in the legal vocabulary.⁵² Holland Michel's is probably the single most useful definition, as it clearly acknowledges that predictability (a) relates to anticipating the future, (b) is concerned with effects rather than functions of a system, and (c) is a matter of degree.

The discourse has also sought to identify different types or forms of predictability. Thus, in the ICRC's view:

There is a ... distinction between predictability in a narrow sense of knowing the process by which the system functions and carries out a task, and predictability in a broad sense of knowing the outcome that will result.⁵³

With regard to the relationship between these two forms of predictability, the ICRC has taken the view that:

predictability in a broad sense [means] knowing the outcome that will result from activating the autonomous weapon system in a particular circumstance. A sub-component of this is predictability in a narrow sense of knowing the process by which the system functions and carries out a specific task or function.⁵⁴

Again, predictability in the narrow sense here might be better described as the interpretability and explainability of the decision-making process of an AI system. Also, the link between the two seems rather tenuous—a system whose technical functioning is completely opaque might nonetheless be reasonably predictable in the broad sense based on extensive testing and evaluation, and/or operational use.

An alternative approach distinguishes predictability in a technical sense and in an operational sense.⁵⁵ From a technical perspective, a system's predictability refers to 'the degree of consistency of its past and current behaviour with its future ones'⁵⁶—in other words, the 'system's ability to execute a task with the same performance that it [has] exhibited [previously]'.⁵⁷ This entails at least two considerations:⁵⁸

- stability—the extent to which the accuracy of the system's outputs remains consistent over time
- generalisability—for machine learning systems, the extent to which the system can handle inputs (data) that deviate from those used for its training and testing.

Technical predictability plainly depends on the system's architecture and features. For example, systems based on deterministic algorithms, such as decision trees, are more predictable than learning systems,⁵⁹ and offline (batch) training models are more predictable than online training models. Put more abstractly, technical predictability depends on the robustness of the system.⁶⁰ Robustness here refers to the ability of the system 'to maintain stable and reliable performance across a broad spectrum of conditions, variations, or challenges, demonstrating resilience and adaptability in the face of uncertainties or unexpected changes'.⁶¹

From an operational perspective, on the other hand, predictability would refer to 'the degree to which an autonomous system's individual actions can be anticipated'⁶² or, perhaps more accurately, 'the degree to which the actions of a system can be anticipated *once it is deployed in a specific environment*'.⁶³

There is broad agreement that all autonomous systems exhibit some degree of inherent operational unpredictability.⁶⁴ As a general matter, following Charles Perrow, it is impossible to anticipate all the interactions within tightly coupled complex systems,⁶⁵ which autonomous military systems made up of sensors, processors and actuators are likely to be. Moreover, 'once deployed, systems may face a context with unforeseen characteristics',⁶⁶ leading to interactions with the environment that their designers and operators could not fully anticipate. Indeed, the very point of devising and deploying autonomous systems is that they should be able handle situations of which the operators lack specific prior knowledge.⁶⁷

With respect to AWS, it has therefore been suggested that the functioning and effect of even a deterministic system ‘will be challenging to predict’,⁶⁸ and that AWS ‘do not function in a broadly predictable fashion, owing to complexity (in design and task) and interaction with a varying environment’.⁶⁹ According to the ICRC:

[A]utonomous weapon systems are unpredictable in a broad sense, because they are triggered by their environment at a time and place unknown to the user who activates them. Moreover, developments in the complexity of software control systems—especially those employing AI and machine learning—may add unpredictability in a narrow sense of the process by which the system functions.⁷⁰

In other words, as noted by Holland Michel:

While technical predictability is solely a function of a system’s performance, operational predictability is just as much a function of the characteristics of the environment and mission for which the system is deployed.⁷¹

Holland Michel goes on to distinguish a ‘third, general, meaning of un/predictability: the degree to which the outcomes or effects of a system’s use can be anticipated’.⁷² This seems to be a distinction without a difference: it is unclear how this ‘third’ form of predictability is any different from operational predictability. The latter appears to precisely capture the ability to anticipate the effects of the system in its operating environment.

Taddeo goes even further. She acknowledges that the predictability problem is ‘multidimensional’ and that the distinction between technical and operational predictability is ‘not really tenable in practice, because technical and operational factors all contribute to determine the behaviour of an AI system’.⁷³ Indeed, it is the combination of inherent technical factors and extrinsic environmental factors that determines the system’s overall predictability.

Therefore, predictability broadly speaking—capturing both technical and operational dimensions—is influenced by myriad factors such as:⁷⁴

- the technical features of the system, including the type of algorithm and the level of computing power
- the complexity of the task or function of the system
- the complexity and characteristics of the operational environment
- the robustness of the system, including the quality of its training data and the rigour of testing
- the system’s capacity to evolve
- the system’s interaction with other systems
- the quality of sensor data supplied to the system

- geographic and temporal scope of the system's operations
- the degree to which the operator understands the way the system operates
- the behaviour of the adversary.

To complicate matters further, '[t]hese variables may change and interact together in complex ways, making it difficult to predict all possible outputs of an AI system and their effects'.⁷⁵ In short, the degree of overall predictability of a system derives from a complex interaction between the technical capabilities of the system, the nature of its task and the features of its operational environment.

From the Predictability Problem to the Predictability Standard

While acknowledging the early advocacy of technologists and scholars, the governance issues of AWS were propelled onto the multilateral agenda through two major lines of effort, which also supplied the early framing of the problem.

On the one hand, in late 2011, the ICRC articulated publicly its concerns about autonomous functions in weapons. In September that year, Jakob Kellenberger, the ICRC's president, gave a keynote speech in which he argued that the deployment of AWS 'would reflect a paradigm shift and a major qualitative change in the conduct of hostilities'.⁷⁶ He went on to acknowledge both the challenges in programming such systems to operate consistently with the law and 'their possible advantages in contributing to greater protection'.⁷⁷ The following month, the ICRC included a brief discussion of these issues in its 'Challenges Report', released ahead of the 31st International Conference of the Red Cross and Red Crescent, the quadrennial meeting of governments and all components of the Red Cross movement.⁷⁸ These early contributions did not, however, expressly mention the predictability problem in relation to AWS.

On the other hand, in April 2013, Christof Heyns, as the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, published an influential report on what he called 'lethal autonomous robots' (LARs).⁷⁹ In that report, he called, *inter alia*, for a national moratorium on LARs and for the development of an international governance framework.⁸⁰ In his substantive findings, Heyns briefly noted the predictability problem in the following terms:

Autonomous systems can function in an open environment, under unstructured and dynamic circumstances. As such their actions (like those of humans) may ultimately be unpredictable, especially in situations as chaotic as armed conflict, and even more so when they interact with other autonomous systems.⁸¹

Perhaps the first comprehensive articulation of this problem in relation to AWS appeared in a background paper that the ICRC produced ahead of an expert meeting that it convened in March 2014.⁸² Questions about unpredictability and its consequences featured prominently throughout the report of that meeting.⁸³ Shortly thereafter, in May 2014, the first informal Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems took place.⁸⁴ In one of its statements, the ICRC clearly connected the predictability problem to compliance with international humanitarian law (IHL):

If weapon systems become more autonomous—with greater ‘freedom’ to determine their operations—they may become less predictable. Unpredictability would raise serious questions with regard to compliance with IHL. For example: What assurance is there that a weapon system will always operate within the law? How can the weapon system be adequately tested and its performance for its intended use verified without predictability in its likely effects? How can the development and deployment of the weapon system be lawful if there is no guarantee that it will perform in accordance with IHL?⁸⁵

At this stage, however, unpredictability remained predominantly an operational concern rather than a normative concept. Experts were grappling extensively with the preliminary question of characterisation—what kinds of systems were under discussion—rather than with the specific legal consequences of unpredictability. However, several experts and delegations stressed the risks associated with the use of AWS in an operational context, which included the lack of predictability but also vulnerability to cyber attacks and difficulties of adaptation to a complex environment.⁸⁶

The 2015 Meeting of Experts largely followed the same pattern: participants continued to discuss unpredictability as a source of operational and legal risk, albeit on a more technical and granular level. Presentations highlighted the challenges posed by unpredictability to legal reviews and legal compliance generally.⁸⁷ Thus, the report of the meeting noted:

Some delegations expressed concern regarding the unpredictability of machine behaviour with self-learning capabilities, in particular when deployed in an unknown or complex environment. The point was made that the lack of determinism and the complexity of the systems would make comprehensive testing difficult, if not impossible. In this context, the importance of limiting the operation of LAWS in time and space was mentioned as a possible solution.⁸⁸

Also, the report noted in cautiously passive voice that the concept of predictability ‘was considered useful for addressing the development of weapon systems, especially those with self-learning capacities’.⁸⁹

In March 2016, the second ICRC-convened expert meeting helped bring further technical understanding and focused in some detail on questions of human control.⁹⁰ Notably, in its background paper, the ICRC articulated a link between predictability and human control. It suggested that there are multiple elements that determine whether human control is meaningful, including ‘predictability of the weapon system in its intended or expected circumstances of use’.⁹¹

The final CCW Meeting of Experts paid significant attention to predictability. Wendell Wallach gave a presentation specifically on that concept, articulating some of the causes of unpredictability and underscoring that predictability is not a binary concept:

Nothing less than a law of physics is absolutely predictable. There are only degrees of predictability, which in theory can be represented as a probability. In the evaluation of weaponry, predictability means that within the task limits for which the system is designed, the anticipated behavior will be realized, yielding the intended result. Nevertheless, an unanticipated event, force, or resistance can alter the behavior of even highly predictable systems. ... Member states may differ on the degree of unpredictability and level of risk they will accept in weapon systems.⁹²

The increased focus on predictability was duly summarised in the meeting report:

The issue of the predictability of autonomous weapons systems was another important aspect of the debate. It was often framed by the notions of risk, reliability and possible differences between human fallibility and malfunctions of machines. Several delegations expressed concern at the prospect of weapons systems that could act unpredictably. It was argued that the control over a system by a military commander is a core capability for the military and determines the value of such systems. A further point was that the possibility of autonomous ‘swarms’ would mean that such systems would be inherently unpredictable.⁹³

Thus, at the time when the informal CCW Meetings of Experts ended and GGE meetings commenced, predictability largely served as a descriptive technical concept rather than a normative standard.

However, in the early work of the GGE, the discussion began to turn to how predictability related to the existing IHL framework. Just to give one example, in a March 2019 statement, the US delegation explained its understanding that the predictability and reliability of AWS are ‘important insofar as they relate to the risk of civilian casualties’ such that an AWS that is ‘unpredictable to the adversary but poses no risk to civilians’ is ‘not a problematic system from the perspective of IHL’.⁹⁴ The US delegation also suggested:

[T]he degree of predictability and reliability ‘required by IHL in order to reduce the risk to civilians must be considered in light of all the circumstances, including the operating environment as well as the other precautions that can be taken to reduce the risk to civilians and civilian objects.’⁹⁵

Summarising the 2020 meetings of the GGE, the chairperson was able to make the broader point that AWS ‘can pose challenges to compliance with IHL rules since predictability and context-specific human decisions are necessary to ensure compliance, especially bearing in mind current technological limitations’.⁹⁶

A pivotal moment in the discussions arrived in 2021 when the ICRC published its detailed position, referenced at the outset of this article, which expressly called upon states to prohibit unpredictable AWS.⁹⁷ This proposal effectively converted (un)predictability from a factor relevant to IHL compliance into a putative freestanding governance standard that could determine a weapon system’s lawfulness. The proposal was plainly influential: predictability subsequently came to feature in proposals for a new CCW protocol submitted by groups of states, as well as, from 2024 onwards, various iterations of the draft elements of an instrument.

Predictability as a Governance Standard

Alignment with Existing Law

The first argument in favour of treating predictability as a governance standard relates to its close alignment with the existing law. Arguably, rules and principles of IHL already imply the need for a weapon system to be predictable, such that articulating predictability as a requirement may be better understood as a clarification of current law rather than as an imposition of an entirely new requirement. This is an important consideration in circumstances where some states resist suggestions that the existing law is somehow insufficient.

Predictability can be grounded in multiple rules currently in force. The prohibition of inherently indiscriminate weapons provides perhaps the strongest anchoring point.⁹⁸ This rule prohibits the employment of means and methods of warfare that are of a nature to strike military objectives and civilians or civilian objects without distinction either because they cannot be directed at a specific military objective or because their effects cannot be limited as required by law.⁹⁹ Traditionally, this prohibition has been read to preclude the use of means and methods that are highly inaccurate or have uncontrollably propagating effects (such as fire or disease).¹⁰⁰ But the prohibition is agnostic as to the precise cause of the indiscriminate effects.¹⁰¹ Thus, an operationally unpredictable AWS, where the operator cannot anticipate whether only military objectives would be engaged by the system, satisfies the definition of an inherently indiscriminate weapon and its use is therefore prohibited.

The rules and principles governing the conduct of hostilities also appear to necessitate predictability. Complying with the principle of distinction, the rule of proportionality and the obligation to take precautionary measures¹⁰² requires the operator to make certain legal judgements during or ahead of an attack. To do this, the operator must be able to reasonably anticipate the effects of the weapon in the circumstances prevailing at the time.

Finally, compliance with the obligation to legally review weapons¹⁰³ presumes that the weapon's effects are predictable. The legal review is in essence a determination ahead of time whether the weapon is capable of lawful use. That determination relies on testing and evaluation carried out before the weapon is fielded. If the weapon does not perform consistently over time, such testing and any legal determination based on it would be meaningless.

Capacity of Being Operationalised

The concept of predictability also seems capable of being articulated by reference to technical and operational criteria. Various characterisations of appropriate human–machine interaction—such as ‘meaningful human control’, ‘appropriate human judgement’ or ‘context-appropriate human judgement and control’—all suffer from indeterminacy, partly owing to the complexity of the socio-technical relationship that they seek to capture. Attempts to define meaningful human control as human control that is not nominal have only kicked the definitional can down the road.

Predictability, on the other hand, seems more readily assessable through robustness, along with appropriate design requirements, as well as testing and evaluation standards. Thus, predictability could be unpacked as a requirement to ensure, through testing across a representative range of operational requirements, that a weapon's performance meets some standard of consistency. Whether a system can be adequately certified as being predictable is at least in principle an empirical question—answerable through technical assessment if appropriate standards are developed.

This is helpful not only from a technical perspective but also from a political one. The notion of appropriate human–machine interaction is subject to considerable disagreement. States disagree as to whether human interaction is a regulatory objective in and of itself, or a means by which some overarching objective (especially compliance with the law) should be achieved. Likewise, states seem to disagree over whether specific technical measures are preconditions for achieving some distinct state of ‘human control’, or whether ‘human control’ is simply the shorthand for the bundle of practical measures that reasonable states should take with a view to achieving compliance with the law. Predictability may carry sufficient technical neutrality to avoid these problems.

Challenges

The notion of predictability comes, however, with its own challenges and limitations. Some of them relate to the very concept while others concern its practical applications.

Open-Endedness

While more concrete than ‘meaningful human control’, ‘appropriate human judgement’ or ‘context-appropriate human judgement and control’, the notion of predictability might still not possess the kind of regulatory precision or technical specificity that would be desirable from a key pillar of a governance framework. The notion immediately raises questions as to what must be predictable, by whom and to what confidence level.

The evolution of the language of the GGE has mitigated part of this problem. Earlier in the debates it was common for participants to refer sweepingly to the predictability (or unpredictability) of the technology or the weapon system. For example, one proposal was to prohibit AWS that ‘operate in a manner that cannot be predicted’ and to take measures to ‘guarantee the weapon system’s ... predictability’.¹⁰⁴ One phrasing more recently considered by the GGE would require states, more specifically, to ‘[e]nsure the effects of LAWS are adequately predictable ... to those responsible for their use’.¹⁰⁵

This phrasing would resolve several issues. First, it would specify that it is the *effects* of the weapon system that need to be predictable, not its operation in some more granular sense. In other words, the focus would be squarely on operational rather than technical predictability. Second, the effect must be predictable *to those responsible* for the weapon’s use—in other words, the operators who are presumably trained and certified to operate the weapon—and not necessarily to other members of the armed forces. Third, the effects must be *adequately* predictable, acknowledging that predictability is not a one-off phenomenon but is rather a matter of degree.

But problems remain. How far must the predictability of effects reach? Does it extend to second- and third-tier effects? It does not help to say here, as one might do in a general IHL analysis, that one must consider *reasonably foreseeable* effects. That would just beg the question. Furthermore, several issues—What level of confidence in predictability is required? How are these levels of confidence articulated? Would AWS with different effects call for different levels of confidence?—would need further examination before predictability could amount to a freestanding standard.

Ex Ante Standards

Considering this uncertainty, the word ‘adequately’ does a lot of heavy lifting in the phrasing mentioned above. But as noted earlier, predictability is at least capable in principle of being expressed in measurable technical and operational criteria. Yet that is not the same as saying that such criteria can be easily developed. Indeed, the opposite appears to be true.

For one thing, there does not seem to be a principled basis for setting particular confidence levels in advance. Any criteria would to a certain degree be arbitrary. Also, the process of developing such criteria would raise challenging questions about the degree to which human behaviour is predictable, and whether any comparison between technical behaviour and human behaviour is appropriate.

Furthermore, any specific criteria are likely to run against the pacing problem—any standard that defines what adequate predictability means for a particular class of AWS may become obsolete rapidly as the technology advances. The behaviour of AWS that is genuinely unpredictable now may become more predictable in the future with new interpretability tools.

Conversely, AWS behaviour deemed adequately predictable today may prove to be inadequate tomorrow with the introduction of more capable and more transparent AI models. For example, during World War II, dropping unguided high-explosive bombs from a high-altitude bomber with the aid of the Norden bombsight was regarded as precision bombing.¹⁰⁶ But this is a far cry from the level of precision that can be achieved with contemporary guided munitions.

Emergent Behaviour and Tactical Unpredictability

A related issue is that in the military context, unlike in the civilian setting, maximum possible predictability of a system is not necessarily desirable. The capacity to surprise an adversary contributes to military effectiveness. Thus, an AWS may be attractive from an operational viewpoint precisely because it behaves in a way that the adversary struggles to anticipate and counteract. As Chris Jenks explains:

There is an advantage when an adversary is not sure how an enemy will conduct operations. Conversely, the more predictable a military force is in combat, the more effective an enemy is in countering. As a result, some amount of emergent behavior (a.k.a. unpredictability) may be desirable.¹⁰⁷

If the behaviour of the AWS is easily predictable to third parties (such as the adversary), those third parties can learn to manipulate the system inputs in order to produce outputs that they desire. That, in turn, may create additional humanitarian risks or reduce the ability to safely operate the system. Edward Hunter Christie and colleagues explain this predicament thus:

[A] machine learning algorithm whose parameters are fully observable by all parties would be explainable and traceable, and thus be compliant with the principle. However, its behaviour will also be—in principle—predictable for all parties, meaning it would violate security needs and offer opportunities to the adversary to predict its actions and gain an unwanted military advantage.

For purposes of military effectiveness, a LAWS should be unpredictable for the adversary, and possibly for the party that operates it, within certain limits. The highest degree of military effectiveness would be achieved if the LAWS could successfully deceive the adversary, while remaining within the bounds of IHL and while also posing no danger to the side that operates it.¹⁰⁸

In short, for the best outcomes, the AWS should be sufficiently predictable to the operator to ensure legal compliance, but sufficiently unpredictable to the adversary to achieve the desired operational effect and prevent manipulation. According to Jenks, '[h]ow much and what kind of emergent behavior [may be desirable] depends on the context in which the system operates but also on a risk calculus'.¹⁰⁹

This problem could be solved from a governance standpoint by defining predictability as relating to the interaction between the AWS and its operator. This is, in fact, something that the GGE had achieved. The first iteration of its elements of an instrument, from July 2024, stipulated in general terms that '[t]o ensure that LAWS can be used in compliance with IHL, their effects must be adequately predictable, reliable, traceable and explainable'.¹¹⁰ The May 2025 version, however, was far more specific: it required states to '[e]nsure [that] the effects of LAWS are adequately predictable, reliable, traceable and explainable *to those responsible for their use*'.¹¹¹ By referring to the persons to whom AWS were supposed to be predictable, the text took a step in the right direction. However, the notion of 'those responsible for their use' is arguably too vague. Technical specificity could be introduced by referring to 'operators' (especially seeing that the GGE refers to the operator in the definition of AWS). Alternatively, using pre-existing legal language, the focal point of predictability could be 'those who plan or decide upon',¹¹² as they bear the obligations to take precautionary measures in the use of the AWS.

Interconnectedness

A further problem is whether predictability constitutes a sufficiently self-contained standard for governance purposes. There are some indications that this might not be the case.

First, it is illuminating how the ICRC, as the most consistent and sophisticated stakeholder on this point, envisages addressing the predictability problem from a regulatory perspective. Fundamentally, the ICRC argues that 'unpredictable AWS should be expressly ruled out'.¹¹³ In the ICRC's view, '[t]his could best be achieved with a prohibition on AWS that are designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained'.¹¹⁴ This suggests that the predictability problem cannot be addressed by simply insisting on a sufficient level of predictability. Rather, the user's ability to predict the effects of a weapon operates in tandem with their ability to understand and explain those effects.

Second, the GGE LAWS has previously tended to use predictability together with reliability. The more recent language contemplated by the GGE has referred to predictability in conjunction with not just predictability but also, similarly to the ICRC approach, traceability and explainability. This, then, pushes to the fore questions about the meaning of those additional desirable attributes, and their overlap or interconnectedness.

By Way of a Conclusion

The foregoing discussion does not definitively resolve the question as to whether, on balance, predictability makes for a good governance standard in a proposed instrument on AWS. However, the paper unpacks some of the conceptual complexities and challenges around the notion itself. These should at least call for some caution. In particular, it seems cavalier to rattle off concepts like predictability, reliability, traceability and explainability as prerequisites of legal compliance. First, at that level of abstraction, these notions are firmly grounded in the discussions about the governance of military AI, not all AWS (which might or might not be enabled by AI). Second, this paper's cataloguing of the various ways in which predictability has been articulated suggests that this concept is not uniformly understood.

With this in mind, a better approach might be to ensure that the text of any legal or policy instrument on AWS reflects the requirement of predictability, even if it does not use this label.¹¹⁵ For example, a requirement to '[e]nsure that the effects of LAWS can be adequately anticipated and controlled by those responsible for their use'¹¹⁶ might capture much of predictability without introducing a term that, by international law standards, remains novel. Admittedly, some of the problems mentioned do not go away—in particular, 'adequately' would do the heavy lifting. But the idea of 'anticipating' the effects of weapons is firmly rooted in international law and can be interpreted in light of longstanding state practice.

Furthermore, the uncertainties around predictability probably go some way to explaining why national policy documents do not utilise the concept. Thus, the Australian policy settings on military AI—though too little, too late¹¹⁷—wisely avoid its inclusion as a discrete principle. But, for better or for worse, predictability has substantial international traction. Therefore, further work to articulate its interrelationship with nationally endorsed AI governance principles might be helpful for multilateral norm development.

About the Author

Dr Rain Liivoja a Professor of Law at the University of Queensland, and a Director of the Asia-Pacific Institute for Law and Security. He has previously been an Expert Advisor to the Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) and a Visiting Legal Fellow with the Department of Foreign Affairs and Trade. His research has covered various legal issues relating military science and technology, military operations and military discipline. He has served on governmental, academic and civil society delegations to multilateral meetings on humanitarian law and arms control, including the GGE LAWS.

ENDNOTES

- 1 I am grateful to Zena Assaad and Neil Renic for the opportunity to present this paper at a workshop held at the Australian National University, and to Jessica Dorsey, Natalia Jevglevskaja and Lauren Sanders for their insightful and helpful comments on earlier drafts. All views and any errors are mine alone.
- 2 International Committee of the Red Cross, *ICRC Position on Autonomous Weapon Systems and Background Paper* (2021), at: www.icrc.org/en/download/file/166330/icrc_position_on_aws_and_background_paper.pdf.
- 3 See e.g. Argentina et al., Working Paper, UN Doc. CCW/GGE.1/2022/WP.4 (8 August 2022), para. 11.
- 4 See e.g. Costa Rica et al., Joint Working Paper (CCW GGE LAWS, June 2021), p. 3, at: <https://documents.unoda.org/wp-content/uploads/2021/06/Costa-Rica-Panama-Peru-the-Philippines-Sierra-Leone-and-Uruguay.pdf>; Argentina et al., Draft Protocol on Autonomous Weapon Systems (Draft Protocol VI), UN Doc. CCW/GGE.1/2023/WP.6 (11 May 2023), s. 1.5.
- 5 On the mandate, see Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Final Report*, UN Doc. CCW/MSP/2023/7 (23 November 2023), para. 20.
- 6 Cf. Group of Governmental Experts on Lethal Autonomous Weapons Systems, Rolling Text (26 July 2024), box III, para. 6, at: <https://perma.cc/T27N-FMUT>; GGE on LAWS, Rolling Text (8 November 2024), box III, para. 6.A, at: <https://perma.cc/7UM6-TAX2>; GGE on LAWS, Rolling Text (12 May 2025), box III, para. 6.A, at: <https://perma.cc/Z8Y9-SD9E>.
- 7 Laura Varella, 'Report of the Meeting of the Group of Governmental Experts on Lethal Autonomous Weapons Systems', *CCW Report* 13, no. 2 (2025): 17, at: <https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2025/gge/reports/CCWR13.2.pdf>.
- 8 GGE on LAWS, Rolling Text (18 December 2025), at: <https://perma.cc/5G4J-JST4>; 'Additional Suggestions from the Chair for Consideration on 5 March 2026 regarding Boxes I–III of the Rolling Text' (4 March 2026), at: <https://perma.cc/8GP6-T8P3>.
- 9 GGE on LAWS, Rolling Text (5 June 2026), para. 15, subpara. C.
- 10 Varella, 'Report of the Meeting of the Group of Governmental Experts on Lethal Autonomous Weapons Systems', p. 17.
- 11 For a comparative overview of the principles, see e.g. Rain Liivoja, 'Principles of Responsible Military Artificial Intelligence and Applicable International Law', GC REALM Expert Policy Note Series (Hague Centre for Strategic Studies, May 2025), at: <https://hcass.nl/wp-content/uploads/2025/05/Liivoja.pdf>.
- 12 Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence: Responsible Use of AI at All Stages of the Technology Lifecycle* (Canberra: Commonwealth of Australia, 2025), at: www.defence.gov.au/sites/default/files/2026-03/Policy-Settings-for-Responsible-Use-of-Artificial-Intelligence-in-Defence-%5BOFFICIAL%5D.pdf.
- 13 See e.g. Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems* (US Government, 2023), at: www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf; cf. International Committee of the Red Cross, Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon Systems (Meeting of Experts on LAWS, 2016), p. 1, at: www.icrc.org/en/download/file/21606/ccw-autonomous-weapons-icrc-april-2016.pdf.
- 14 GGE on LAWS, Rolling Text (5 June 2026), para. 1, subpara. A.
- 15 International Committee of the Red Cross, *Views of the ICRC on Autonomous Weapon Systems*, p. 1.
- 16 GGE on LAWS, Rolling Text (5 June 2026), para. 1 ('a combination of one or more weapons and functionally integrated technological components').
- 17 *Ibid.*, para. 1, subpara. B.
- 18 *Ibid.*, para. 1, subpara. C.

- 19 See e.g. US remarks summarised in Laura Varella, 'Report on the Virtual Informal Consultation of the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS)', *CCW Report* 13, no. 3 (2025): 6, at: www.reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2025/gge/reports/CCWR13.3.pdf.
- 20 See e.g. BJ Copeland, 'Artificial Intelligence', *Encyclopaedia Britannica*, 20 March 2026, at: www.britannica.com/technology/artificial-intelligence; Office for Artificial Intelligence, *National AI Strategy* (London: UK Government, 2021), at: www.gov.uk/government/publications/national-ai-strategy.
- 21 *Information Technology—Artificial Intelligence—Artificial Intelligence Concepts and Terminology*, International Standard ISO/IEC 22989:2022(E), July 2022, para. 3.1.3–3.1.4, at: www.iso.org/standard/74296.html.
- 22 Dave Bergmann, 'What Is Machine Learning?', *Think*, 18 August 2025, at: www.ibm.com/think/topics/machine-learning.
- 23 For an overview, see e.g. Global Commission on Responsible AI in the Military Domain, *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (HCSS, 2025), pp. 11–19, at: <https://hcss.nl/report/gc-reaim-responsible-by-design-strategic-guidance-report>.
- 24 See e.g. Frank Sauer, 'Stepping Back from the Brink: Why Multilateral Regulation of Autonomy in Weapons Systems Is Difficult, yet Imperative and Feasible', *International Review of the Red Cross* 102, no. 913 (2020): 241, at: <https://doi.org/10.1017/S1816383120000466>; yet Imperative and Feasible, *International Review of the Red Cross* 102, no. 913 (2020).
- 25 Vincent Boulanin and Maaïke Verbruggen, *Mapping the Development of Autonomy in Weapon Systems* (SIPRI, 2017), p. 17, at: www.sipri.org/publications/2017/other-publications/mapping-development-autonomy-weapon-systems.
- 26 Sauer, 'Stepping Back from the Brink', p. 241.
- 27 See e.g. Ishmael Bhila, 'Putting Algorithmic Bias on Top of the Agenda in the Discussions on Autonomous Weapons Systems', *Digital War* 5, no. 3 (2024): 201–212, at: <https://doi.org/10.1057/s42984-024-00094-z>.
- 28 Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata', *Ethics and Information Technology* 6, no. 3 (2004): 175, at: <https://doi.org/10.1007/s10676-004-3422-1>; the manufacturer/operator of a machine is held (morally and legally).
- 29 *Ibid.*, p. 175.
- 30 Robert Sparrow, 'Killer Robots', *Journal of Applied Philosophy* 24, no. 1 (2007): 65, 70, at: <https://doi.org/10.1111/j.1468-5930.2007.00346.x>.
- 31 Mariarosaria Taddeo and Alexander Blanchard, 'Accepting Moral Responsibility for the Actions of Autonomous Weapons Systems: A Moral Gambit', *Philosophy and Technology* 35, no. 3 (2022): 6, at: <https://doi.org/10.1007/s13347-022-00571-x>.
- 32 Norbert Wiener, 'Some Moral and Technical Consequences of Automation', *Science* 131, no. 3410 (1960): 1355–1358, at: <https://doi.org/10.1126/science.131.3410.1355>.
- 33 *Ibid.*, p. 1355.
- 34 I am grateful to Jessica Dorsey for drawing my attention to this point.
- 35 Arthur L Samuel, 'Some Moral and Technical Consequences of Automation: A Refutation', *Science* 132, no. 3429 (1960): 741–742, at: <https://doi.org/10.1126/science.132.3429.741>.
- 36 *Ibid.*, p. 741.
- 37 Indeed, 'the predictability problem is central to any analysis of ethical implications of AI in defence'. Mariarosaria Taddeo, *The Ethics of Artificial Intelligence in Defence* (Oxford University Press, 2024), p. 3, at: <https://doi.org/10.1093/oso/9780197745441.001.0001>.
- 38 Roman V Yampolskiy, 'Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent', *Journal of Artificial Intelligence and Consciousness* 7, no. 1 (2020): 109, at: <https://doi.org/10.1142/S2705078520500034>.
- 39 Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI* (UNIDIR, 2020), p. 1, at: <https://unidir.org/publication/the-black-box-unlocked>.
- 40 Yampolskiy, 'Unpredictability of AI', p. 110.

- 41 Tobias Vestner and Altea Rossi, 'Legal Reviews of War Algorithms', *International Law Studies* 97, no. 1 (2021): 535, at: <https://digital-commons.usnwc.edu/ils/vol97/iss1/26>.
- 42 See Vienna Convention on the Law of Treaties, 23 May 1969, 1155 UNTS 331, art. 31.
- 43 *Oxford English Dictionary*, 'predictability (n.)', last modified July 2023, at: <https://doi.org/10.1093/OED/2868482080>.
- 44 *Oxford English Dictionary*, 'predictable (adj.)', last modified March 2026, at: <https://doi.org/10.1093/OED/7210877849>.
- 45 *Oxford English Dictionary*, 'predict (v.)', last modified December 2025, para. 1.a, at: <https://doi.org/10.1093/OED/1009911946>.
- 46 See e.g., Neil Davison, 'A Legal Perspective: Autonomous Weapon Systems Under International Humanitarian Law', in *Perspectives on Lethal Autonomous Weapon Systems*, UNODA Occasional Paper No. 30 (UNODA, 2017), p. 10, at: www.icrc.org/sites/default/files/document/file_list/autonomous_weapon_systems_under_international_humanitarian_law.pdf.
- 47 International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (Geneva, 2019), p. 10, at: https://www.icrc.org/en/download/file/102852/autonomy_artificial_intelligence_and_robotics.pdf; relying on Davison, 'A Legal Perspective', p. 10.
- 48 Andrea Iacona, 'Knowledge of Future Contingents', *Philosophical Studies* 179, no. 2 (2022): 447, at: <https://doi.org/10.1007/s11098-021-01666-5>.
- 49 Rome Statute of the International Criminal Court, 17 July 1998, 2187 UNTS 90, art. 30(3).
- 50 Taddeo and Blanchard, 'Accepting Moral Responsibility', p. 6 (*italics omitted*).
- 51 Holland Michel, *The Black Box, Unlocked*, p. 5.
- 52 See e.g. Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 8 June 1977, 1125 UNTS 3 (Additional Protocol I), art. 51(5)(b), art. 57(2)(a)(iii) and art. 85(3)(c).
- 53 International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics*, p. 27 (*emphases removed*).
- 54 *Ibid.*, p. 10.
- 55 Holland Michel, *The Black Box, Unlocked*, p. 5.
- 56 Taddeo and Blanchard, 'Accepting Moral Responsibility', p. 7.
- 57 Holland Michel, *The Black Box, Unlocked*, p. 7.
- 58 *Ibid.*, p. 5; Taddeo and Blanchard, 'Accepting Moral Responsibility', p. 7.
- 59 International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics*, p. 13; Taddeo and Blanchard, 'Accepting Moral Responsibility', p. 7.
- 60 Taddeo and Blanchard, 'Accepting Moral Responsibility', p. 7.
- 61 Houssein Ben Braiek and Foutse Khomh, 'Machine Learning Robustness: A Primer', in Marco Lorenzi and Maria A Zuluaga (eds), *Trustworthy AI in Medical Imaging* (Academic Press, 2025), p. 37, at: <https://doi.org/10.1016/B978-0-44-323761-4.00012-2>.
- 62 Holland Michel, *The Black Box, Unlocked*, p. 5.
- 63 Taddeo, *The Ethics of Artificial Intelligence in Defence*, p. 5 (*emphasis added*).
- 64 Holland Michel, *The Black Box, Unlocked*, p. 5.
- 65 See generally Charles Perrow, *Normal Accidents: Living with High-Risk Technologies*, updated edition (Princeton University Press, 2011), at: <https://doi.org/10.1515/9781400828494>.
- 66 Taddeo and Blanchard, 'Accepting Moral Responsibility', p. 7.
- 67 Holland Michel, *The Black Box, Unlocked*, pp. 5, 24, n. 8.
- 68 Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (SIPRI and ICRC, 2020), p. 7, at: https://www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf.
- 69 International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics*, p. 12.
- 70 *Ibid.*, p. 11.
- 71 Holland Michel, *The Black Box, Unlocked*, p. 5.

- 72 Ibid., p. 6.
- 73 Taddeo, *The Ethics of Artificial Intelligence in Defence*, p. 6.
- 74 Holland Michel, *The Black Box, Unlocked*, pp. 6–7; Taddeo, *The Ethics of Artificial Intelligence in Defence*, pp. 5–6.
- 75 Taddeo, *The Ethics of Artificial Intelligence in Defence*, pp. 5–6.
- 76 Jakob Kellenberger, 'Keynote Address—International Humanitarian Law and New Weapon Technologies: 34th Round Table on Current Issues of International Humanitarian Law, San Remo, 8–10 September 2011', *International Review of the Red Cross* 94, no. 886 (2012): 812, at: <https://doi.org/10.1017/S1816383112000793>.
- 77 Ibid., pp. 812–813.
- 78 International Committee of the Red Cross, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: 31st International Conference of the Red Cross and Red Crescent, 28 November–1 December 2011*, Doc. 31IC/11/5.1.2 (ICRC, 2011), pp. 39–40.
- 79 Christof Heyns, *Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions*, UN Doc A/HRC/23/47 (2013).
- 80 Ibid., paras 113–126.
- 81 Ibid., para. 42.
- 82 International Committee of the Red Cross, *Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects (Expert Meeting, Geneva, Switzerland, 26–28 March 2014)* (ICRC, 2014), pp. 57–94, at: <https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf>.
- 83 Ibid.
- 84 *Report of the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, UN Doc CCW/MSP/2014/3 (2014).
- 85 International Committee of the Red Cross, 'Statement of the International Committee of the Red Cross', Meeting of Experts on Lethal Autonomous Weapons Systems, Geneva, 13 May 2014, p. 3, at: [https://uno-da-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_\(2014\)/ICRC_MX_LAWS_2014.pdf](https://uno-da-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2014)/ICRC_MX_LAWS_2014.pdf).
- 86 *Report of the 2014 Informal Meeting of Experts on LAWS*, para. 36.
- 87 *Report of the 2015 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, UN Doc CCW/MSP/2015/3 (2015), para. 51.
- 88 Ibid., para. 34.
- 89 Ibid., para. 42.
- 90 International Committee of the Red Cross, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons (Expert Meeting, Versoix, 15–16 March 2016)* (ICRC, 2016), p. 83.
- 91 Ibid.
- 92 Wendell Wallach, 'Predictability and Lethal Autonomous Weapons Systems (LAWS)', Meeting of Experts on LAWS, Geneva, 12 April 2016, at: <https://perma.cc/5JMG-HCR7>.
- 93 *Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS)*, UN Doc CCW/CONF.V/2 (2016), para. 40.
- 94 United States of America, 'Consideration of the Human Element in the Use of Lethal Force', GGE LAWS, Geneva, 26 March 2019, at: <https://geneva.usmission.gov/2019/03/27/convention-on-certain-conventional-weapons-consideration-of-the-human-element-in-the-use-of-lethal-force>.
- 95 Ibid.
- 96 *Chairperson's Summary*, UN Doc. CCW/GGE.1/2020/WP.7 (2021), para. 6.
- 97 International Committee of the Red Cross, *ICRC Position on Autonomous Weapon Systems and Background Paper*, p. 2.
- 98 See e.g. *ibid.*, p. 7.
- 99 See Additional Protocol I, art. 51(4)(b)–(c).

- 100 Claude Pilloud and Jean de Preux, 'Protocol I—Article 51. Protection of the Civilian Population', in Yves Sandoz et al. (eds), *Commentary on the Additional Protocols to the Geneva Conventions* (ICRC, 1987), at: <https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-35/commentary/1987>.
- 101 Aleks Kajander, Rain Liivoja and Maarja Naagel, 'Cybersecurity of Weapon Systems: International Law Requirements and Technical Standards', *Journal of Cybersecurity* 11, no. 1 (2025), at: <https://doi.org/10.1093/cybsec/tyaf017>.
- 102 See especially Additional Protocol I art. 48, art. 51(5)(b) and art. 57.
- 103 See *ibid.*, art. 36.
- 104 Argentina et al., *Draft Protocol on Autonomous Weapon Systems (Draft Protocol VI)*, UN Doc CCW/GGE.1/2023/WP.6 (2023), arts 3, 4(1)(5).
- 105 GGE on LAWS, Rolling Text (12 May 2025), box III, para. 6.A.
- 106 Raymond Patrick O'Mara, *Rise of the War Machines: The Birth of Precision Bombing in World War II* (Naval Institute Press, 2022).
- 107 Chris Jenks, 'Responsible AI Symposium—The AI Ethics Principle of Responsibility and LOAC', *Articles of War*, 21 December 2022, at: <https://lieber.westpoint.edu/ai-ethics-principle-responsibility-loac>.
- 108 Edward Hunter Christie, Amy Ertan, Laurynas Adomaitis and Matthias Klaus, 'Regulating Lethal Autonomous Weapon Systems: Exploring the Challenges of Explainability and Traceability', *AI and Ethics* 4, no. 2 (2024): 229–245, at: <https://doi.org/10.1007/s43681-023-00261-0>.
- 109 Jenks, 'Responsible AI Symposium—The AI Ethics Principle of Responsibility and LOAC'.
- 110 GGE on LAWS, Rolling Text (26 July 2024), box III, para. 6.
- 111 GGE on LAWS, Rolling Text (12 May 2025), box III, para. 6.A (emphasis added).
- 112 See Additional Protocol I, art. 57(2)(a).
- 113 International Committee of the Red Cross, *ICRC Position on Autonomous Weapon Systems and Background Paper*, p. 8.
- 114 *Ibid.*, p. 8.
- 115 Asia-Pacific Institute for Law and Security, 'Statement on Box III, Paragraph 6', GGE LAWS, 3 September 2025, at: <https://apils.org/2025/09/03/statement-on-box-iii-paragraph-6>.
- 116 GGE on LAWS, Rolling Text (18 December 2025), box III, para 7(B).
- 117 For a more nuanced and eloquent critique, see Netta Goussac and Zena Assaad, 'Australia's New Military AI Policy Comes at a Crucial Time. The Challenge Is Turning It into Practice', *The Conversation*, 25 March 2026, at: <https://doi.org/10.64628/AA.5ctjas7h7>.

/ ARCHITECTING TRUST: A 'TRUST BY DESIGN PLUS' FRAMEWORK FOR MILITARY AI

Branka Marijan

Introduction

Artificial intelligence (AI) is increasingly embedded across military functions, including intelligence analysis, logistics optimisation, targeting support and strike.¹ These technologies promise enhanced speed, precision and decision advantage, while simultaneously introducing new forms of uncertainty, opacity and risk. The increasing reliance on AI in time-sensitive and high-stakes environments raises fundamental questions about how human decision-makers understand, evaluate and rely on machine-generated outputs. Indeed, recent reporting on the February 2026 US attack on Iran, reportedly involving AI-enabled decision-support systems across stages ranging from target identification to legal review and strike launch, has been interpreted as signalling a shift towards increasingly compressed decision-making timelines, raising concerns that human judgement may become progressively marginalised in the use of force.²

Within this evolving landscape, trust has emerged as both an operational requirement and a strategic concern. For military organisations, personnel must trust AI-enabled systems sufficiently to be willing to employ them in operational settings, as low levels of trust may significantly constrain the circumstances in which such capabilities are used and, consequently, their operational effectiveness. At the same time, policymakers must ensure that these systems operate in ways consistent with legal and ethical obligations. Beyond the organisational level, states must also interpret and assess each other's development and deployment of AI-enabled military capabilities, often under conditions of uncertainty and strategic competition.

Trustworthy AI in the military domain has therefore emerged as a shared concern among military practitioners and policymakers. According to a 2025 European Defence Agency white paper:

Trustworthy AI systems in the military domain are designed to inspire confidence among commanders, soldiers, and policymakers by ensuring that AI technologies operate predictably, securely, and ethically in complex and dynamic environments.³

In parallel, international discussions on the responsible application of AI in the military domain, including debates on lethal autonomous weapons systems, have increasingly highlighted (dis)trust among states as a central issue.

Despite the growing attention to trust and distrust, the topic remains underexplored in interdisciplinary policy and scholarly literature on military AI.⁴ One of the few empirical studies in relation to trust and military attitudes toward AI systems finds that:

service members' trust in partnering with AI during strategic-level deliberations is based on a tightly calibrated set of considerations, including technical specifications, perceived effectiveness, and international oversight.⁵

This latter point on international oversight is particularly significant, as broader policy discussions are at times described as disconnected from the practical considerations that shape military decision-making.

At the same time, the expanding technical literature on trust and trustworthiness in AI systems has tended to overlook the interstate dynamics that shape how AI is developed, deployed and perceived in international security contexts. Nevertheless, insights from this literature have influenced regulatory discussions on military AI, particularly as technical experts have been increasingly engaged in these processes. Addressing these gaps requires a more integrated understanding of how trust is constructed, operationalised and sustained across multiple levels of analysis in the military AI domain.

This article addresses some of these gaps by advancing three interrelated arguments. First, trust in military AI should be understood as a multi-level construct spanning system, organisational and interstate dimensions. Second, trust is dynamic and must be continuously calibrated across the lifecycle of AI systems, rather than treated as a static design outcome. Third, trust is politically situated and should therefore be calibrated conservatively through demonstrated performance and continuous oversight, rather than maximised as an end in itself. This reflects concerns that claims of 'trustworthy AI' may be instrumentalised to legitimise or accelerate the adoption of military AI systems without sufficient scrutiny of their limitations and risks.

Building on these insights, this paper argues that the trust by design (TbD) framework, as conceptualised by Emmanuel A Merchán-Cruz et al. in the context of industrial human–robot collaboration,⁶ offers a useful foundation for structuring trust in military AI systems when adapted to account for the socio-technical complexities of international security. By embedding principles such as transparency, human agency and robustness into system design, an expanded 'trust by design plus' (TbD+) framework may provide states with more credible and potentially verifiable signals of restraint, while reinforcing human control in the use of force. Rather than seeking to maximise trust, it emphasises the importance of calibrated trust grounded in demonstrable system performance, institutional oversight and adherence to shared norms. Trust, in this sense, is not an input to be engineered but a property that must be continuously earned and validated. It also extends beyond the technical system itself and must account for how users interact with both the system and the operational environment in which it is deployed.⁷ This, in turn, requires closer integration between system design and development processes and the ways in which military organisations introduce, govern and employ these capabilities in service. Therefore, the TbD+ framework does not substitute for political and strategic judgement regarding the use of AI-enabled systems in the first place.

Additionally, technical and governance design approaches may still be insufficient to address deeper trust deficits in international security, and notably between major powers such as the United States and China. Accordingly, the paper also identifies a set of complementary confidence-building measures aimed at reducing risks of miscalculation, including commitments to shared principles on human control in the use of force, incident notification protocols, information-sharing on testing and evaluation standards, and capacity-building initiatives within multilateral institutions. Ultimately, this layered approach and expanded understanding of trust by design in the military domain will also require additional research and study, particularly as policies, global dynamics and real-time deployments of technologies continue to develop.

On Trust and Control in Military AI

Trust is often described as a 'slippery' term, with multiple meanings and definitions across different fields of study.⁸ For the purposes of this discussion, rather than engaging with the extensive literature on the many interpretations of trust, it is more useful to focus on scholarship that examines how trust emerges in relation to new technologies such as AI, particularly within the military domain.

But before discussing some of this literature, it is important to acknowledge that much of the debate on military AI, particularly in discussions on autonomous weapons, has focused on human control. Tim McFarland's examination of trust and control in the

military use of AI highlights a divergence in how these concepts have been treated that is critical to the points made in this paper.⁹ Discussions of control have been central in regulatory and policy debates, while AI developers have placed greater emphasis on trust. This dynamic may be partially explained by institutional and commercial incentives. Trust represents a significantly lower threshold for industry to achieve, whereas designing systems that genuinely enable ‘control by design’ often requires substantial state involvement early and throughout the development lifecycle, increasing costs and, in many cases, running up against commercial barriers and procurement rules that limit state capacity to participate in co-design or effectively ‘pick winners’. As a result, the two paradigms have largely evolved in parallel.

In international discussions on lethal autonomous weapons systems, for example, states have emphasised the capacity of humans to exercise control over weapons systems and have outlined measures to ensure that such control is maintained. As McFarland observes, ‘The idea of relying on ‘trust’ in the system to behave as desired is effectively absent from regulatory discussions’.¹⁰ At the same time, McFarland notes that both trust and control are shaping the application of military. McFarland asks, ‘Might there be a complementary relationship between correctly calibrated trust and “meaningful human control”?’¹¹

As debates on lethal autonomous weapons and responsible military AI, particularly through successive Responsible AI in the Military Domain (REAIM) summits, have evolved, it has become increasingly evident that many militaries are seeking to ensure both appropriately calibrated trust in AI systems and the preservation of human judgement.¹² The relationship between trust and human control has become increasingly explicit, including during discussions among military practitioners and academics at the third REAIM Summit, held in Spain in 2026. While the summit’s outcome document uses the term trust only once, it repeatedly emphasises the governance mechanisms that underpin justified trust, including transparency; testing, evaluation, validation and verification (TEVV); accountability, reliability; human oversight; and confidence-building measures. Rather than centring the language of ‘trustworthy AI’, the document places greater emphasis on ensuring that AI-enabled military systems are governable, accountable and verifiable. A similar evolution is evident in the 2026 Policy Settings for Responsible Use of Artificial Intelligence in Defence, released by the Australian Defence Force and Department of Defence, which explicitly link human judgement with ‘trusted autonomy’.¹³ Recognising the importance of these concepts, however, is only a starting point. States must still determine how AI-enabled capabilities can be designed, assessed, tested and introduced in ways that meaningfully support human judgement and foster appropriately calibrated trust in operational practice.

Taken together, these developments suggest a shift from the earlier separation identified by McFarland, in which trust and control were treated as largely distinct paradigms. Instead, they are increasingly understood as interdependent: effective human control

depends on appropriately calibrated trust in system performance, while such trust, in turn, is shaped by the presence of meaningful human oversight and the ability to exercise judgement. That said, this evolving understanding remains conceptually underdeveloped. Informal discussions at REAIM summits suggest that trust is often assumed rather than clearly defined, with the term used in broad and sometimes ambiguous ways. In particular, equating trust with system reliability alone risks overlooking how AI systems shape human decision-making beyond technical performance alone.¹⁴ A system may be relied upon because it performs well in certain conditions, but trust involves a broader judgement that includes ethical, institutional and contextual factors. In this regard, scholarship on human agency in military AI systems is particularly important.¹⁵ Whether explicitly acknowledged in formal discussions or not, achieving appropriately calibrated levels of trust will be essential for defining and regulating human–machine interaction requirements in practice.

Trust, Trustworthiness and the Politics of Trust

In fact, how control and agency is maintained is going to be shaped in part by the technical architecture of trust. Indeed, trustworthiness has been extensively studied in more technical AI literature, and technical experts have been present in shaping international discussion on lethal autonomous weapons and responsible military AI. In the context of AI systems, trust has often been understood in terms of reliance: whether users choose to depend on system outputs in decision-making processes. In military settings, understanding trust primarily in terms of reliance is appealing, as personnel often have limited choice in whether to use particular systems, either because those are the capabilities available to them or because operational demands create an immediate need for their use. However, this narrow conception has been increasingly challenged by research in human–AI interaction.

Siddharth Mehrotra et al. provide a useful overview of the complexity of different approaches to ensuring appropriate trust in technical literature. From their review of the literature, they note that the approaches focus on confidence scores, explanations, cues, alarms, warning signals and uncertainty communication.¹⁶ Some of the approaches seem particularly relevant for militaries applications of AI. For example, Q Vera Liao and S Shyam Sundar argue (emphasis in original):

Trustworthiness of a technology is not inherently established but communicated through *trustworthiness cues*, which are embedded in interface features, documentation, and other modes of information, such as speech acts for conversational AI.¹⁷

These cues are especially relevant in safety-critical contexts as they shape the way in which decisions could be made in time-compressed environments such as conflict zones.

Importantly, Mehrotra et al. point to the fact that recent research on human–AI interaction further complicates the notion of ‘appropriate’ or calibrated trust.¹⁸ Some of the literature has implicitly equated calibration with system accuracy, distinguishing over-trust from under-trust based on whether AI outputs are correct.¹⁹ However, this approach has been increasingly challenged. Trust cannot be reduced to correctness alone; rather, it is shaped by a range of factors, including perceived reliability, transparency, interpretability, prior interactions and users’ cognitive and emotional engagement with the system.²⁰

This multi-dimensional understanding of trust has important implications for military AI. In operational contexts characterised by uncertainty, time pressure and adversarial interference, trust is not static but dynamic. It must adapt to changing system performance, environmental conditions and task requirements. A system that performs reliably in one context may degrade in another, requiring operators to continuously recalibrate their expectations and reliance. Yet existing literature provides limited guidance on how such recalibration occurs over time, particularly in long-term or high-stakes deployments.²¹

At the 2024 REAIM Summit in Seoul, experts such as Missy Cummings noted the need for certification and recertification of systems in order to address concerns about using the technology beyond the contexts it has been trained and tested for. Vincent Boulanin and Dustin A Lewis similarly emphasise the importance of ‘responsible reliance’ in ensuring accountability for the use of military AI systems.²² They argue that states must:

devise and maintain a framework that ensures that natural persons involved in the development and use of a military AI tool can responsibly rely on the tool’s technical aspects, on the other members of the group and on the state itself.²³

As a result, responsible reliance must be maintained across the lifecycle of AI-enabled capabilities and embedded within the broader organisational and political structures that govern their development, deployment and use.

Moreover, many military decisions are not binary but involve probabilistic assessments, trade-offs and incomplete information. In such cases, identifying over- or under-trust becomes significantly more complex, as there are no clear benchmarks for ‘correct’ reliance. Indeed, in many military contexts it may be neither practical nor desirable to establish these fixed benchmarks. Instead, decisions about whether and how to rely on a particular system are likely to depend on contextual understanding, operational judgement and applicable legal requirements. This underscores the need to move beyond simplified models of trust calibration and towards more context-sensitive approaches that account for the specific characteristics of tasks and decision environments.

These insights reinforce the argument advanced in this article: trust in military AI cannot be treated as a fixed or purely technical attribute. Rather, it is a dynamic, multi-dimensional relationship that must be continuously shaped through system design, organisational practice and broader governance frameworks.

Beyond organisational and technical considerations, trust in military AI must also be understood as a political construct. As Sian Troath (2024) argues, the growing emphasis on 'trustworthy' and 'ethical' AI in military contexts is not neutral but can function to facilitate the development and adoption of AI-enabled and autonomous systems while deflecting attention from more fundamental political questions. Troath notes that these include who such systems are used against, under what conditions and for what strategic purposes.²⁴ In this sense, trust operates not only as a relational property between users and systems but also as part of a broader discourse that can legitimise particular forms of technological development and military practice.²⁵ The coupling of 'trust' with 'ethics' risks narrowing debate to questions of how systems function, rather than whether their use is justified in specific contexts.

For military AI, this raises an important analytical and normative challenge. Efforts to design for trust must avoid becoming mechanisms that facilitate uncritical adoption or obscure political accountability. Instead, trust should be understood as contingent, limited and subject to ongoing scrutiny. The TbD+ framework is therefore positioned as a means not of maximising trust but of constraining and calibrating it. By embedding mechanisms that expose system limitations, reinforce human agency and support oversight, the framework seeks to ensure that trust remains proportionate to system capabilities and embedded within broader processes of political and legal accountability.

Core Principles of Trust by Design

In order to fully build out the TbD+ framework, it is necessary to understand the conceptualisation by scholars focused on the issue in human–robot industrial research. The framework is particularly useful to military AI applications as it provides a structured approach to embedding trustworthiness into system design.

At its core, the TbD framework conceived by Merchán-Cruz et al. for industrial human–robot collaboration identifies several interrelated design logics that are particularly salient when adapted to military AI. First, it foregrounds human agency and empowerment, ensuring that AI systems augment rather than replace human judgement. This includes preserving the capacity for operators to interrogate, adjust and override system outputs, particularly in high-stakes contexts involving the use of force. Second, it emphasises transparency and interpretability not as abstract principles but as functional requirements that allow users to understand system behaviour, assess uncertainty and anticipate

outcomes. Third, it highlights the importance of robustness and safety, including fail-safe mechanisms, predictable performance and the visibility of system limitations under varying conditions.

In addition, the framework incorporates accountability and traceability, ensuring that decisions and actions can be reconstructed and evaluated, as well as data governance and fairness, addressing the ethical implications of data use and potential bias. Finally, it underscores the role of user integration and training, recognising that trust is shaped through repeated interaction, organisational context and user competence, rather than design alone. Yet, within military contexts the challenge is not simply articulating such principles but determining how they are operationalised across the lifecycle of capabilities. Human agency, transparency and accountability must therefore be translated into concrete practices of testing, validation, oversight, doctrine and operational governance if they are to shape the use of AI systems in practice rather than remaining aspirational design goals.

Trust by Design Plus (TbD+)

The trust by design plus (TbD+) framework builds on the above work while addressing its limitations in military contexts. While frameworks such as the European Commission's Ethics Guidelines for Trustworthy AI and related defence policy documents emphasise principles such as transparency, accountability and human oversight, they often treat trustworthiness as a set of desirable system attributes rather than as a dynamic and relational condition. TbD+ extends these approaches by integrating technical, organisational and political dimensions of trust, recognising that trust in military AI is constructed across multiple levels of interaction and shaped by broader strategic contexts.

The framework is guided by principles across three levels: systems level, organisational level and interstate level. At the system level, the framework focuses on the design features and technical characteristics that condition human interaction with AI systems. Drawing on the TbD literature, this includes human agency and empowerment, transparency and interpretability, and robustness and safety. In military contexts, these elements take on heightened significance: systems must not only perform reliably but also make their limitations visible, support operator understanding under time pressure and enable meaningful human intervention in high-stakes decisions. Crucially, these features do not produce trust directly. Rather, they shape the conditions under which users form expectations, assess uncertainty and adjust reliance.

At the organisational level, TbD+ extends beyond system design to consider the institutional context in which AI systems are deployed. Here trust is shaped by doctrine, training, oversight mechanisms and professional norms that structure human-machine

interaction. Organisational practices determine how responsibility is allocated, how uncertainty is managed and how human agency is preserved or eroded in operational settings. This dimension is particularly important in mitigating well-documented risks such as automation bias, over-reliance and the diffusion of responsibility in complex socio-technical systems.

The organisational dimension raises a broader set of institutional questions that military organisations must address if trust in AI-enabled capabilities is to be appropriately calibrated in practice. These include how operational responsibility and accountability are distributed across chains of command; how professional military judgement is maintained under conditions of increasing automation and compressed decision-making timelines; and how doctrine, procurement and capability development processes shape the integration of AI into operational environments.²⁶ TbD+ also raises questions regarding how organisations identify where human agency remains essential, how operators are trained to understand system limitations and uncertainty and how institutional incentives may encourage either excessive reliance on or resistance to AI-enabled systems.

Additional organisational challenges include ensuring interoperability and shared understandings of trust across different services, commands and allied forces; integrating civilian and military oversight mechanisms; managing adaptation as systems evolve over time; and determining how lessons from operational use, failure and near misses are incorporated back into training, testing and governance processes. In this sense, trust at the organisational level is a matter not simply of user confidence but of institutional capacity to govern human–machine interaction responsibly across the lifecycle of AI-enabled capabilities.

At the interstate level, the framework situates trust within broader strategic and political dynamics. In this domain, claims of 'trustworthy AI' function as signals that are interpreted by other states under conditions of uncertainty. Realising this dimension of TbD+ in practice requires far deeper integration between international policy discussions, capability development processes and the operational governance of AI-enabled systems than currently exists. This includes linking emerging norms and political commitments to concrete practices of testing, evaluation, doctrine, procurement and operational oversight across the lifecycle of capabilities. Given current geopolitical tensions, this may be the most difficult element of TbD+ to operationalise, yet it is also increasingly important both for enabling interoperability and shared understandings among allies and for reducing risks of miscalculation and escalation with adversarial states.

Trust is not directly observable but is inferred through practices such as transparency measures, information-sharing and adherence to emerging norms. At the same time, these signals are inherently ambiguous and subject to contestation. As a result, trust in military AI is not only a technical or organisational matter, but also a political one,

shaped by strategic incentives, competing narratives and the risk of misinterpretation. Importantly, these principles across the three levels are not intended to generate trust in any straightforward sense. Rather, they function as constraints that define the conditions under which trust may be considered warranted. This distinction is critical in light of concerns that trust can be instrumentalised as a means of facilitating the adoption of AI systems without adequately addressing their risks or broader implications.

TbD+ therefore reframes trust not as an output of system design but as a condition of use that must be governed across time, context and interaction. It emphasises that trust cannot be calibrated solely through system accuracy or static design features. Instead, trust must be continuously evaluated and, where necessary, recalibrated based on system performance, user experience, organisational practice and the evolving strategic environment. In this sense, TbD+ aligns with a broader shift away from technocratic understandings of trust towards an approach that recognises its dynamic, relational and inherently political character.

A Lifecycle Approach: Operationalising Trust by Design Plus

For the TbD+ framework to be operationally useful, it must be implemented alongside a lifecycle approach to military AI governance. Lifecycle governance has become increasingly prominent in discussions on responsible military AI and autonomous weapons because it recognises that issues of human control, judgement and agency cannot be addressed at a single point in the development process but must instead be considered throughout the evolution of an AI-enabled capability.²⁷ Reflecting this shift, the Institute of Electrical and Electronics Engineers (IEEE) 2024 White Paper on Military Autonomous and Intelligent Systems (AIS) identifies nine interconnected stages spanning the capability lifecycle, from pre-development planning and system requirements through research and development, procurement, TEVV, human–system integration, operational employment, and ultimately review, reuse, or retirement.²⁸

This lifecycle perspective reinforces a central premise of TbD+: trust is not an inherent property of an AI system; nor is it established at a single point in time. Rather, appropriately calibrated trust is constructed, assessed, maintained and, where necessary, recalibrated throughout the lifecycle as systems are designed, tested, integrated, deployed, operated, adapted and ultimately retired. During the early stages of capability development, this includes identifying legal, ethical, and operational risks, defining intended use and incorporating user-centred design principles. During TEVV, it involves evaluating not only technical performance but also human–machine interaction, operator understanding, and the calibration of reliance under realistic operational conditions. Deployment and operational use require appropriate training, doctrine, leadership and

continuous monitoring of both system performance and human behaviour, while review and retirement provide opportunities to capture lessons learned and feed them back into future capability development and governance.

This lifecycle perspective also recognises that trust is dynamic rather than static. It may increase, decrease or fluctuate depending on system performance, operational context and organisational practice. Systems that perform reliably in one environment may perform differently in another, requiring operators and organisations to continually reassess and recalibrate their reliance. Appropriately calibrated trust therefore depends not only on technical design but also on institutional arrangements, including doctrine, oversight, professional education and mechanisms for organisational learning.

Table 1 outlines selected considerations for implementing the TbD+ framework across the IEEE military AI lifecycle. Given the diversity of state requirements, operational contexts and approaches to military AI governance, it is intended as an illustrative guide rather than a comprehensive or prescriptive checklist.

Table 1. Implementing TbD+ across the IEEE military AI lifecycle

IEEE military AI lifecycle stage	System level	Organisational level	Interstate level
Requirements	Define intended use, human decision points and acceptable limits of autonomy.	Establish operational need, allocate responsibilities and identify risks to human judgement.	Align with international commitments and interoperability objectives.
Design and development	Build transparency, interpretability, robustness and fail-safe mechanisms into the system.	Set procurement criteria, involve users in design reviews, and plan training and change management.	Design for compatibility with allies and consider export controls and responsible transfer.
TEVV	Test technical performance, safety and limitations; evaluate human-machine interaction and calibration.	Conduct human-machine teaming exercises, assess operator performance and document lessons.	Use agreed testing approaches; share non-sensitive methodologies and results where appropriate.
Integration	Ensure usable interfaces, human override and integration with existing systems.	Deliver training; update doctrine, roles and authorities.	Ensure interoperability and shared understandings with allied forces.
Deployment	Confirm system readiness and operational safeguards.	Certify operators; establish rules of engagement and command responsibilities.	Communicate capability intent and appropriate use; support confidence-building.
Operations	Monitor performance in context; communicate uncertainty and support operator calibration.	Monitor use, collect feedback, manage risks (e.g. automation bias, over-reliance) and adapt training.	Engage in information sharing where appropriate; maintain strategic signalling and stability.
Maintenance and updates	Update models and data; re-test and validate changes.	Manage configuration control; update doctrine and training; ensure oversight of updates.	Notify allies/partners of significant changes where appropriate.
Incident review and learning	Conduct technical root-cause analysis; identify contributing factors.	Conduct accountability review; capture lessons from incidents and near misses.	Share lessons and best practices; contribute to norm development where appropriate.
Retirement/disposal	Securely decommission systems and preserve audit trails.	Capture knowledge and documentation; manage workforce transition.	Provide transparency on retirement/disposal where appropriate.
Across all stages	Apply human agency, transparency, robustness, safety and accountability principles.	Maintain governance, oversight, training and institutional learning.	Support international law, interoperability and confidence-building measures.

Importantly, TbD+ does not resolve the broader political question of whether and under what conditions military AI should be deployed. Rather, it operates within these constraints, recognising that technical and governance interventions cannot substitute for political judgement. At the same time, the framework is intended to help inform such judgements by providing a structured approach for assessing where human agency is essential, how trust can be appropriately calibrated and what forms of testing, oversight and organisational governance are required across the lifecycle of AI-enabled capabilities. In this sense, TbD+ seeks not to legitimise the expansion of AI-enabled warfare but to govern and constrain the conditions under which such systems may be considered appropriate for use.

Trust and Confidence Building at the Interstate Level

Trust in military AI extends beyond technical and organisational dimensions to encompass interstate dynamics and broader political considerations. As Troath observes, the logic of 'trust but verify' captures a defining feature of international security and arms control: trust must be grounded in observable practices rather than rhetorical commitments.²⁹ Dialogue and information-sharing are therefore not neutral confidence-building exercises but mechanisms through which states seek to produce, interpret and contest evidence about each other's behaviour.

This dynamic reflects core insights from international relations theory, where trust is not assumed but inferred through signalling, interpretation and repeated interaction under conditions of uncertainty, as outlined by Andrew Kydd.³⁰ From this perspective, claims of 'trustworthy AI' are not simply technical descriptors but strategic interventions that can shape expectations, frame acceptable uses of technology and potentially narrow political debate. As such, they must be approached with caution: rather than serving as straightforward indicators of responsible behaviour, they are themselves subject to interpretation, contestation and, at times, instrumentalisation.

Building on this understanding, confidence-building measures remain an important, if inherently limited, mechanism for managing uncertainty and mitigating risks of miscalculation in the development and deployment of military AI.³¹ Such measures may include mutual declarations in which states affirm shared principles, such as the maintenance of human control in the use of force; the establishment of incident notification protocols or dedicated communication channels to clarify AI-related errors or unintended behaviour before they escalate; and the development of joint or comparable approaches to testing and evaluation standards, aimed at reducing the risks associated with system brittleness or unpredictability.³² At the same time, declarations alone are insufficient to establish trust. Their credibility ultimately depends on whether subsequent

state practice consistently reflects and reinforces the commitments being made. In this sense, confidence-building measures should be understood less as guarantees of responsible behaviour than as mechanisms through which states generate evidence, signal intentions and enable ongoing assessment of whether stated commitments are borne out in practice.

In addition, capacity-sharing initiatives may support the development of responsible governance frameworks across a wider range of states, helping to reduce asymmetries that could otherwise contribute to instability. Michael Horowitz notes that at least some of these measures might be achievable. Horowitz points out:

The military AI areas most likely to generate agreements, on average, will be those where international cooperation can reduce accidents, inadvertent escalation and miscalculation involving military applications of AI, since these are in the interest of all or nearly all countries.³³

However, consistent with the analysis above, these measures should not be understood as generating trust in any straightforward sense. Rather, they function as mechanisms for structuring expectations, enabling interpretation and constraining uncertainty within an inherently competitive and politically contested environment. As a result, their effectiveness will depend not only on their design but on the extent to which they are embedded in sustained practices of engagement and subject to ongoing scrutiny. It is also worth recognising that poorly designed or implemented confidence-building measures may unintentionally undermine trust by creating ambiguity, unrealistic expectations or strategic suspicion among states.

Conclusion

This article makes a conceptual contribution by reframing trust in military AI as neither a purely technical property nor a stable relational outcome but as a dynamic and politically situated process. Existing approaches often treat trust as something to be engineered through system design or measured through user reliance. By contrast, this article argues that an architecture of trust is multi-dimensional, context dependent and continuously evolving across human-machine interaction, organisational practice and interstate dynamics. Building on this insight, the proposed TbD+ framework shifts the focus from maximising trust to constraining and calibrating it. In doing so, it brings together technical, organisational and political perspectives demonstrating that trust in military AI must be governed as an ongoing condition of use rather than assumed as a prerequisite for adoption. Additional research and study is needed to examine the proposed framework.

At the same time, the political dimensions of trust extend beyond interstate dynamics alone. The broader 'trust ecosystem' surrounding military AI includes relationships between industry, military organisations, governments, academia, civil society and the public, as well as differing layers of trust within military institutions themselves, including between services and between military and civilian elements. These relationships shape how AI-enabled capabilities are developed, interpreted, adopted and contested across the lifecycle of use. Future research should therefore examine how the TbD+ framework and its lifecycle approach may assist both researchers and practitioners in navigating this wider trust ecosystem, including how trust is negotiated, maintained and challenged across institutional, societal and strategic contexts over time.

In addition, the maintenance of human agency is critical to ensuring appropriately calibrated trust, avoiding both over-reliance on AI-enabled systems and susceptibility to automation bias or related cognitive effects in AI-supported decision-making. This requires militaries to first identify where and how human agency is currently exercised, and where it remains essential, before effective safeguards can be designed to preserve it. Existing command-and-control structures often implicitly assume full human participation across intelligence, decision-making and teaming functions, an assumption that may be increasingly challenged as AI systems take on more operational and cognitive tasks. This is particularly relevant given the growing role of AI decision-support systems and recognition of the impact on human roles and responsibilities in human-machine teaming.³⁴

Ultimately, the challenge is to determine when, to what extent and under what conditions specific AI-enabled systems should be trusted in particular operational contexts. By reframing trust as a dynamic, constrained and politically situated relationship, this article contributes to a growing body of work that seeks to move beyond technocratic optimism towards a more grounded understanding of AI in military contexts.³⁵ In doing so, it highlights that the governance of military AI is a question not only of capability but of judgement, restraint and responsibility. As contemporary cases of AI-enabled warfare are demonstrating, there is a pressing need to consider the purposes and consequences of using AI-enabled systems.

About the Author

Dr Branka Marijan is a CIGI senior fellow and a senior researcher at Project Ploughshares. She is a lecturer in the Master of Global Affairs program at the Munk School of Global Affairs and Public Policy at the University of Toronto. At Ploughshares, Branka leads research on the military and security implications of emerging technologies. Her work examines concerns regarding the development of autonomous weapons systems and the impact of artificial intelligence and robotics on security provision. Her research interests include trends in warfare, civilian protection, use of drones and civil-military relations. She holds a Ph.D. from the Balsillie School of International Affairs with a specialization in conflict and security. She has conducted research on post-conflict societies and published academic articles and reports on the impacts of conflict on civilians and diverse issues of security governance, including security sector reform. Branka closely follows United Nations disarmament efforts and attends international and national consultations and conferences. She is a board member of the Peace and Conflict Studies Association of Canada and a research fellow at the Kindred Credit Union Centre for Peace Advancement at the University of Waterloo.

ENDNOTES

- 1 Michael C Horowitz, *Artificial Intelligence, the Future of War and International Politics*, CIGI Papers No. 345 (CIGI, 2026).
- 2 Robert Booth and Dan Milmo, 'Iran War Heralds Era of AI-Powered Bombing Quicker than "Speed of Thought"', *The Guardian*, 3 March 2026.
- 3 European Defence Agency, *Trustworthiness for AI in Defence: Developing Responsible, Ethical, and Trustworthy AI Systems for European Defence* (European Defence Agency, 2025), pp. 12–13).
- 4 Roff, Heather M., and David Danks. 2018. "'Trust but Verify': The Difficulty of Trusting Autonomous Weapons Systems." *Journal of Military Ethics* 17 (1): 2–20. doi:10.1080/15027570.2018.1481907.; Tim McFarland, 'Reconciling Trust and Control in the Military Use of Artificial Intelligence', *International Journal of Law and Information Technology* 30, no. 4 (2022): 472–483; Sian Troath, 'Trusting Technology to Wage War: The Politics of Trust and Ethics in the Development of Robotics, Autonomous Systems, and Artificial Intelligence', *Critical Military Studies* 11, no. 1 (2025): 59–77; Branka Marijan, *Trust by Design? AI in Military Applications*, CIGI Policy Brief no. 218 (CIGI, 2025); Paul Lushenko, 'AI, Trust, and the War Room: Evidence from a Conjoint Experiment in the US Military', *Cambridge Forum on AI: Law and Governance* 1: e52 (2025).
- 5 Lushenko, 'AI, Trust, and the War Room', p. 2.
- 6 Emmanuel A Merchán-Cruz, Ioseb Gabelaia, Mihails Savrasovs, Mark F Hansen, Shwe Soe, Ricardo G Rodriguez- Cañizo and Gerardo Aragón-Camarasa, 'Trust by Design: An Ethical Framework for Collaborative Intelligence Systems in Industry 5.0', *Electronics* 14, no. 10 (2025): 1952.
- 7 Ingvild Bode and Katherine Chandler, 'Re-Thinking Human–Machine Interaction and the Governance of AI in the Military Domain', *Nature Machine Intelligence* 8 (2026).
- 8 S Krüger and C Wilson, 'The Problem with Trust: On the Discursive Commodification of Trust in AI', *AI & Society* 38 (2023): 1753–1761; Roy Lindelauf and Herwin Meerveld, 'Building Trust in Military AI Starts with Opening the Black Box', *War on the Rocks*, at: <https://warontherocks.com/2025/08/buildingtrust-in-military-ai-starts-with-opening-the-black-box>.
- 9 McFarland, 'Reconciling Trust and Control in the Military Use of Artificial Intelligence'.
- 10 Ibid., p. 477.
- 11 Ibid., p. 480
- 12 Lindelauf and Meerveld, 'Building Trust in Military AI Starts with Opening the Black Box'.
- 13 Australian Defence Force and Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence: Responsible Use of AI at All Stages of the Technology Lifecycle* (Canberra: Commonwealth of Australia, 2026), p. 5.
- 14 Alon Jacovi, Ana Marasović, Tim Miller and Yoav Goldberg, 'Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2021), pp. 624–35.
- 15 Ingvild Bode, *Human-Machine Interaction and Human Agency in the Military Domain*, CIGI Policy Brief no. 193 (Waterloo ON: CIGI, 2025).
- 16 Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker and Myrthe L Tielman, 'A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges', *ACM Journal on Responsible Computing* 1, no. 4, article 26 (2024), p. 4.
- 17 Q Vera Liao and S Shyam Sundar, 'Designing for Responsible Trust in AI Systems: A Communication Perspective', in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2022), pp. 1257–1268.
- 18 Mehrotra et al., 'A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction'.
- 19 Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- 20 Liao and Sundar, 'Designing for Responsible Trust in AI Systems'.
- 21 Mehrotra et al., 'A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction'.

- 22 Vincent Boulanin and Dustin A Lewis, 'Responsible Reliance Concerning Development and Use of AI in the Military Domain', *Ethics and Information Technology* 25, no. 8 (2023).
- 23 Ibid.
- 24 Troath, 'Trusting Technology to Wage War'.
- 25 Daniel Møller Ølgaard, 'The New Technopolitics of War: (Re)imagining Agency and Authority in Military Affairs', *Global Policy* 16, no. 3 (2025).
- 26 Netta Goussac, 'Responsible Behaviour in Military AI Starts with Responsible Procurement', *SIPRI* (website), at: www.sipri.org/commentary/essay/2025/military-ai-responsible-procurement.
- 27 Zena Assaad and Jessica Dorsey, 'Designing Lawful Military AI: Technical and Legal Reflections on Decision-Support and Autonomous Weapon Systems', *Perry World House* (website), at: <https://perryworldhouse.upenn.edu/news-and-insight/designing-lawful-military-ai-technical-and-legal-reflections-on-decision-support-and-autonomous-weapon-systems>.
- 28 IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems, *A Framework for Human Decision Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (New York NY: IEEE SA, 2024).
- 29 Troath, 'Trusting Technology to Wage War'.
- 30 Kydd, Andrew H. 2005. *Trust and Mistrust in International Relations*. Princeton, NJ: Princeton University Press.
- 31 Michael C Horowitz, Lauren Kahn and Casey Mahoney, 'The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?', *Orbis* 64, no. 4 (2020): 528–543.
- 32 Ioana Puscas, *Confidence-Building Measures for Artificial Intelligence: A Framing Paper* (Geneva: United Nations Institute for Disarmament Research, 2022).
- 33 Horowitz, *Artificial Intelligence, the Future of War and International Politics*, p. 2.
- 34 Jessica Dorsey and Marta Bo. 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human (e) Dimension', *International Law Studies* 106 (2025); Zena Assaad, 'A Risk-Based Trust Framework for Assuring the Humans in Human-Machine Teaming', *TAS '24: Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, no. 3 (New York NY: Association for Computing Machinery, 2024): pp. 1–9.
- 35 Neil Renic and Elke Schwarz, 'Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing', *Ethics & International Affairs* 37, no. 3 (2023): 321–343.

/ MILITARY AI SYSTEMS AND THE DESIGN OF LIFECYCLE- SENSITIVE GOVERNANCE FRAMEWORKS

Rob McLaughlin and Conor McLaughlin

Introduction

As Natalie Nunn recently observed in the context of using the Convention on Certain Conventional Weapons (CCW) process and experience as a guide for autonomous weapon systems (AWS) governance development:

If one reads the CCW closely, a pattern emerges. Where States drew clear, technically legible regulatory lines early in weapons development, as with blinding laser weapons, the law shaped development pathways. Where States relied on incremental restriction for a weapon already entrenched in their arsenals, as was the case with landmines, regulation became longer, harder, and ultimately vulnerable to geopolitical reversal. And most importantly, where definitional choices left loopholes, such as the case with incendiary weapons, the law struggled to keep up with how weapons were actually used.¹

Nevertheless, as a general rule, on those few occasions where the law has tried to get ahead of technology—to pre-empt through prohibition—this has often led to unintended consequences, or has subsequently been considered ‘off with the fairies’. An example of the former is flattening and expanding rounds, insofar as these rounds have evolved since their early prohibition.² Such rounds are today regularly used by policing and counterterrorism forces (including military forces) because of their ‘stopping power’, while remaining prohibited to military forces in operations governed by the law of armed conflict (LOAC).³ An example of the latter is the ENMOD Convention,⁴ which was negotiated in the 1970s in relation to suspicions and fears around the use of technology to weaponise the environment. These fears have not eventuated to date, although it is possible that ENMOD will become more acutely relevant in the near term if current reports about the Russian Poseidon nuclear torpedo (which could create a radioactive tsunami) are

accurate.⁵ It is also instructive that the ENMOD idea was in part a ‘look over there’ Cold War era distraction to keep non-nuclear weapon states away from seeking to engage in treaty negotiations on nuclear weapons. As Emily Crawford has noted, the ENMOD Convention was in some respects ‘attributable to a desire, on the part of both the US and the USSR, to be seen as acting on disarmament issues, if not actually acting on disarmament issues’.⁶ Indeed, it is interesting to ask whether the current mode of engagement by some states with existing negotiations around AI in a military context is likewise an attempt to be seen to be acting, while in fact hampering action on this issue.

There is no reason to suppose that the current legal debate around regulation of military AI, or AI with military uses, will not run very similar risks. The preoccupation with defining and classifying, rather than ‘regulating backwards from desired effects and state practice’,⁷ for example, is an indicator of this challenge. Another indicator is continuing debate as to the adequacy of the current rules to regulate AWS generally. On one hand, for example, the Secretary-General’s 2025 report in response to United Nations General Assembly (UNGA) Resolution 79/239⁸ observes that there is a clear concern as to the application of existing law, including international humanitarian law (IHL):

The affirmation by the General Assembly in its resolution 79/239 that international law, including the Charter of the United Nations, international humanitarian law and international human rights law, applies throughout the life cycle of AI is an important baseline. However, important questions on how the law applies remain to be resolved.⁹

Similarly, the 2025 Draft Articles on Autonomous Weapon Systems¹⁰ proposes that the idea of new articles on (in this case) lethal autonomous weapon systems (LAWS) IHL regulation is not about new ‘hard law’, but rather is designed to support proper and coherent interpretation of existing specific IHL rules.

On the other hand, it is the equally clear view of many stakeholders that the existing governance frameworks for AI in a military context are already inadequate to the technology, and the solution is to develop a fresh governance framework based on new treaty instruments and rules.¹¹ This option requires new ‘hard law’—that is, legally binding rules. Hard law is generally considered to be rules contained within the sources of law set out in article 38(1) of the Statute of the International Court of Justice (ICJ):

- a. international conventions, whether general or particular, establishing rules expressly recognized by the contesting states;
- b. international custom, as evidence of a general practice accepted as law;
- c. the general principles of law recognized by civilized nations;

d. subject to the provisions of Article 59 [ICJ decisions bind only the parties to it], judicial decisions and the teachings of the most highly qualified publicists of the various nations, as subsidiary means for the determination of rules of law.¹²

However, designing a governance framework that requires new hard law at its core will face significant practical challenges. Chief among these is that the ambition to negotiate, agree on, and then implement a new treaty on AI in a military context is an almost impossible ask in the current geopolitical context.¹³ This is perhaps most readily evident in the glacial pace of negotiations and debate in international fora, and the fact that it has taken at least a decade to build the political will to bring the topic even to the UNGA:

While the General Assembly has adopted multiple resolutions related to the peaceful uses of AI, it only began considering the subject of AI in the military domain at its seventy-ninth session 2024. Previously, such discussions had taken place in several plurilateral groups outside the UN.¹⁴

The consequence, in terms of hard law, is that we need to use what we already have, which means that we need to regulate by leveraging and interlocking existing hard law general principles and specific prohibitions with new soft law instruments.

Soft law, in contrast to hard law, is rules that are politically/operationally and interpretively influential but not legally binding. Daniel Thurer defines soft law as follows:

Soft law is used in legal literature to describe principles, rules, and standards governing international relations which are not considered to stem from one of the sources of international law enumerated in Art. 38(1) ICJ Statute.¹⁵

However, because soft law is less formalised, not governed by the Vienna Convention on the Law of Treaties (VCLT) and treaty-making processes, and not legally binding, it can be developed and applied in myriad ways and formats, and through a much wider array of fora and institutions. States are (currently) moderately more willing to engage in soft law projects than in hard law projects,¹⁶ so using soft law mechanisms to stitch together existing hard law rules in a lifecycle governance regime for AI in a military context is a significantly more realistic (albeit still restrained) ambition than negotiating a new hard law regulatory regime specifically targeted at AI in a military context.

Consequently, this paper argues that a new hard law enterprise is not (currently) the right approach to take with respect to designing AI lifecycle governance frameworks for military contexts. Rather, the paper proposes three guardrails that could perhaps usefully inform the development—or rather evolution—of a governance framework for AI in military contexts that is based in existing rules, mechanisms and processes. The paper then briefly notes some of the key components of a lifecycle governance framework that are already in place and the need to identify the point of diminishing returns by which a new regime would need to be developed.

Three Guardrails to Inform the Design of a Useful and (for the Time Being) Enduring Governance Framework

The basic proposition underpinning the following guardrails is that it is unrealistic, in the current context, to assert that a single governance mechanism, or a single metric, might be developed to cover the life cycle of AI in a military context. Certainly, LOAC and weapons law are essential components of any such regime, but they cannot provide the totality of that regime because they cannot cover a complete life cycle. Nor can LOAC and weapons law cover the penumbra that surrounds technology regulation more broadly, such as export control. That is, the development of an effective lifecycle governance regime for AI in military contexts will also need to learn from, and employ elements of, a wide range of weapons- and technology-adjacent regulatory schemes such as non-proliferation law,¹⁷ the Arms Trade Treaty,¹⁸ international standards regulatory systems, incident investigatory regimes and so on.

An effective lifecycle governance regime will of necessity need to be piecemeal, leveraging different regulatory mechanisms, concepts, and options for different aspects of the life cycle—design, precursor management, trade and technology transfer, development and acquisition, test and evaluation, employment, updating, and decommissioning. It is difficult to envisage, for example, what a singularly focused and all-encompassing regulatory treaty would look like in respect of lifecycle governance of AI in a military context. It is more realistic to envisage such a regime as a Lego construction of interlocking or touching mechanisms and sub-regimes regulating different aspects of the technology's use in military contexts in a coherent, linear—albeit sectoral—manner.

An example of relative (albeit currently stressed) success in such a Lego regime approach is nuclear non-proliferation, created through an interlocking set of chapeau treaties (e.g., the 1968 Nuclear Non-Proliferation Treaty¹⁹ and the 1996 Comprehensive Nuclear-Test-Ban Treaty²⁰), implementation treaties (such as the Comprehensive Safeguards Agreements between states and the International Atomic Energy Agency (IAEA)²¹), sectoral enforcement treaties (such as the International Convention for the Suppression of Acts of Nuclear Terrorism²² and the Protocol of 2005 to the 1988 Convention for the Suppression of Unlawful Acts Against the Safety of Maritime Navigation²³), liability treaties (the 1977 Vienna Convention on Civil Liability for Nuclear Damage²⁴), security and transport treaties and guidelines (including the 1979 Convention on the Physical Protection of Nuclear Material and its 2005 Protocol²⁵ and the regularly updated IAEA Regulations for the Safe Transport of Radioactive Material²⁶), and other soft law instruments designed to fill gaps and join the seams of hard law regimes (such as UNGA resolutions²⁷ and codes of conduct²⁸). With this in mind, we propose three guardrails to help inform the development of a lifecycle governance regime for AI in military contexts.

Don't Regulate Prospective Technology; Regulate Process

As noted above, attempts to comprehensively regulate specific new technologies before their scope and utility is adequately understood (flattening and expanding rounds), or prospectively before they are developed (ENMOD) carry risk. One risk is unintended consequences, such as inconsistency or even illogicality in operational limitations. Another risk is the regulatory regime's inability to evolve as the technology evolves because the prohibition was pre-emptively comprehensive—particularly considering the intention for AI to learn, adapt and innovate.²⁹ Another risk is that precursor comprehensive regulation of a specific military technology can also roadmap for states how to avoid regulation by developing similar weapons effects in a particular way so as to avoid the application of that regulatory instrument.

An alternative approach is to regulate the process by which technologies—including anticipated technologies—are to be developed, acquired, assessed and employed. This approach is point-in-time neutral, and can generally be achieved through one or more interlocking process-based regulatory approaches, three of which are highly relevant to AI: the 'system of control', principles-based regulation, and assurance regulation. However, it is essential to note at the outset that the solution is unlikely to be found solely in any one of these approaches. Rather, the solution is likely to be found in an amalgam and synchronisation of these various approaches. That will in itself present an organisational and regulatory challenge, but—again—it is a challenge of sequencing and linking rather than of ab initio creation.

The System of Control Approach

In its 2019 non-paper submitted to the United Nations Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS), Australia proposed a 'system of control' approach to LAWS regulation.³⁰ In this paper, the concept of 'system of control' was defined as 'an incremental, layered approach to applying control, covering all aspects of a weapon system from design through to engagement'.³¹ That is, a system of control approach employs already existing legal processes and rule sets in an interlocking, mutually reinforcing and lifecycle-focused manner to provide a holistic governance regime. Components of this system of control include research and development regulations, article 36 reviews (including as an ongoing obligation³²), targeting processes, command responsibility assessments, and end-of-lifecycle decommissioning requirements. Certainly the military uses of AI are broader than just LAWS, but there is no obvious reason why a similarly lifecycle-focused amalgam of processes would not be achievable for this broader capability. Indeed, an approach that covers all military applications of AI would, from a regulatory coherence perspective, be far preferable to a stove-piped set of parallel but separate regulatory schemes.

Principles-Based Regulation

UK policy states:

We will work closely with allies and partners to build consensus, promote a common vision for the safe, responsible and ethical use of these technologies globally, and push for compliance with International Humanitarian Law.³³

This approach differs from the system of control approach in that it focuses on principles rather than specific system-embedded rules. This approach also tends to be more ecumenical in terms of the sources of and concepts for a regulatory system, precisely because it is based on principles rather than rules. For example, the UK approach to principles-based regulation of AI and LAWS, as noted above, includes reference to safety, responsibility and ethics, as well as IHL, as sources for these principles. However, the challenge with this approach, as compared to the system of control approach, is that it would be ineffective without a precursor (and likely contentious) negotiation as to the fundamental principles which will underpin the governance regime, whereas the system of control approach tends to leverage a matrix of already existing rules and rule sets. And as the Stockholm International Peace Research Institute has noted, the diversification and fragmentation of fora and the breadth of the debate, and the trade-offs that will be required to crystallise outcomes, will continue to challenge translation of principles into practical guidelines.³⁴

Assurance-Based Regulation

The assurance approach focuses particularly on the substantive, measurable content of an 'assurance' outcome. That is, the assurance approach is animated by the requirements of generating adequate levels of trust, rather than by a focus on direct, compliance-indicating, specific hard law obligations. An example is the US Department of Defense/War's 'AI ethical principles', which assess five factors within its assurance metric:

Responsible. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

Equitable. The Department will take deliberate steps to minimize unintended bias in AI capabilities.

Traceable. The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.

Reliable. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

Governable. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.³⁵

An assurance-focused governance regime, consequently, is not *prima facie* concerned with potentially hard to pin down granular definitions of the technology itself, or the details of how or why it is used, but rather with how contextually trustworthy it is assessed to be. It differs from the system of control approach in that it is governed by an outcome metric—trust—rather than a list of specific rule compliances in the context of a relatively linear set of metrics. Assurance also differs from principles-based governance in that it is targeted at a perhaps more explicable and unitary measure—trust—rather than at a matrix assessment of potentially differently weighted and situationally variable principles. However, as with the principles-based approach, the development of an agreed assurance metric (such as trust) again presupposes a precursor agreement as to how that metric is defined and measured.

Conclusion

Use of process-based regulatory systems for as yet incompletely evolved and understood technologies (including AI technologies) that are capable of use in a military context is one method by which some of the risks of unintended limitation, 'rainbow unicorn' regulation, and providing a roadmap for avoiding regulation—which can result from a focus on the specific technology rather than on the process of regulation—might be avoided.

Don't Regulate to the Nightmare (Terminator); Regulate to What Is Actually in the 'Pipeline'

It is difficult to argue against the proposition that *The Terminator* is currently used as a touchpoint for debate around how to regulate AI in a military context. As one group of scholars have recently asserted of this linkage:

The notoriety of LAWS is not surprising given the predatory, apocalyptic light in which they are commonly cast. In the media and elsewhere, LAWS are frequently identified with the killer robots of *The Terminator* movie.³⁶

Or, perhaps more viscerally, as Ford has observed:

Who, at this point, *doesn't* approach a discussion of lethal autonomous weapons systems without thinking, even inadvertently, of the villainous 'Skynet' ... and the robot-assassins it dispatches against the noble but hard-pressed remnants of humanity in Hollywood's Terminator franchise?³⁷

Menadue has similarly argued that 'the Terminator is ... considered the poster boy for any debate on lethal autonomous weapons'.³⁸

The problem with trying to regulate to the nightmare as the first priority rather than focusing on what is actually and realistically anticipated—that is, basing regulatory assumptions on possible future, but as yet impractical, technology, rather than on near-term anticipated technological evolution—is precisely that it runs the risk of not dealing with what is in train. It is the 'now' and the 'soon' which ought to be the initial focus of an effective and implementable regulatory system.³⁹ The problem with focusing on 'sci-fi' technology that does not yet exist, as John Lewis argued as far back as 2015, is that 'the time to act is now' arguments in relation to these sci-fi techno-fears tend to 'exaggerate the danger posed ... and underappreciate regulation's potential to respond to the threats'.⁴⁰ This is not to say that the Terminator nightmare should be ignored; rather it is to say that achieving a stable and scalable regulatory scheme for the now and the soon should provide experience that can be deployed to address this more distant concern if and as it evolves.

Don't Focus on One Mechanism or a Single Novel Metric; Regulate by Multiple Mechanisms and by Reference to Already Existing and Defined Metrics and Contexts

The proposition behind this guardrail is that it is impractical to seek to create and define a new regulatory silver bullet—such as meaningful human control, a concept that has preoccupied the CCW discussions for more than decade thus far.⁴¹ Rather, a less fraught option may be to choose metrics or limiters that are more stable and enduring—such as domain, place in the 'kill chain', and/or consequence—even if this requires stitching together multiple processes rather than the development of a singular, standalone, all-encompassing process.

Meaningful human control has been defined as:

Elements of human control [which are] types of control measures necessary to ensure legal compliance, ethical acceptability and operational utility ... any kind of measure that acts as a necessary constraint on AWS and that could form the basis of internationally agreed standards or rules.⁴²

It is the assumed lack of such meaningful, humanly defined and recognisably human controls that is the essence of the meaningful human control (MHC) debate. But the problem with focusing so heavily on MHC as the silver bullet metric for governance is that as a concept, it is difficult to define and operationally assess, and has little textual support in existing law. As Gerald Mako has observed of MHC:

[T]he Gaza war already pointed to MHC's serious limitations. Israeli Defense Force operators were reportedly approving AI-generated targets in an average of 20 seconds with minimal scrutiny, leading to a 10% error rate and thousands of civilian deaths. This underscores that while MHC may very well be effective in smaller operations, in large-scale and/or drawn-out conflicts where the speed and volume of AI decisions overwhelm operators, it is unlikely that rushed oversight will prevent IHL violations.⁴³

Perhaps a better approach might be to start by regulating via an already established or known concept, metric or paradigm. This approach would offer a discrete but scalable start point, and there are a variety of options that immediately present.

The first option is to iteratively regulate by domain. For example, a scheme could start with regulation of AI in naval warfare contexts, where the issues of identification, distinction and proportionality are to some extent easier to manage than situations on land because of the platform-based targeting rules.⁴⁴ Once the 'easier' issues are resolved, the scheme could then step up to the next domain (air, cyber, space?) before then progressing to what is undoubtedly the most complex domain in this regard: land.

A second option is to regulate by place in the kill chain. This could involve beginning at one end of the spectrum and then iteratively building—or reducing—from that point. For example, a regulatory regime could begin with AI that does supply or logistics but has little or no role in the kill chain; then move to enablers that get successively closer to the end effect, such as sensor managers or decision-support managers; then reach AI that directly delivers a destructive effect. The lessons learned by first regulating further from the lethal effect could then inform the evolution of the regulatory scheme as it leans further into complexity while approaching the delivery of lethal effects.

A third option is to regulate by consequence. This approach would prioritise what it is that using the AI 'causes'. This is more than just an analysis of risk simpliciter, as it includes quantitative and qualitative assessments of the nature of the harm. This is in some ways the EU approach in respect of prohibited practices, as encapsulated in the 2024 EU Artificial Intelligence Act.⁴⁵ For example, in a military context, if the AI kills or destroys, LOAC already has clearly applicable rules which are well embodied and embedded in the targeting process, and these rules already have a built-in balancing mechanism: the principle of proportionality. If the AI *enables* a consequence, LOAC already has criteria

for assessing what first and second order effects, or reverberating consequences, are relevant to the assessment,⁴⁶ and also has rules about what acts or roles or conduct have the consequence of characterising that ‘thing’ as direct participation in hostilities and thus describing proximity to destructive consequences.⁴⁷ The point is that it may be more efficient, effective and sensible to use a known paradigm, concept or metric to underpin how we consider, assess and sequence the regulation of AI operations, rather than trying to impose a single, new, ill-defined and legally indeterminate metric as the regulatory silver bullet.

Conclusion

The three guardrails above are offered as suggestions for how we approach the development of a (currently) enduring and useful governance regime for AI in a military context. However, equally important to any regulatory development enterprise is the identification of necessary inputs. In this regard, there are three key inputs to or components of such a regime which we can say are already in place (knowns), but two key inputs or components necessary for the regime which would need to be identified (unknowns). Additionally, there are a number of more variable components which are quantifiable only in point-in-time assessments, such as reaching the tipping point of political will to develop a governance regime. Given their fluidity, uncertain variables such as political will not be further addressed; however, the three key knowns and the two key unknowns are more capable of general description and will be briefly outlined below.

Existing Components

The three already existing components of a lifecycle governance regime for AI in a military context are the existing law, the existence of recognised mechanisms for evolutive interpretation of existing law, and a generally permissive approach to soft law development.

The Substantive Law/Rules (and the Expertise to Interpret and Apply Them)

The Australian government has told the UN’s GGE on LAWS that:

Australia advocates for building a shared understanding of how existing IHL applies to LAWS before pursuing any new legal instrument. Australia is fully committed to building this shared understanding through the ongoing discussions in the GGE. Accordingly, we do not support the creation of a parallel process on LAWS that would be detrimental to our collective efforts in the GGE.⁴⁸

That is, among the myriad existing rule sets that directly (LOAC, weapons law, command responsibility) or indirectly (counter-proliferation, technology transfers, international standards) govern AI that can be used in military contexts, there is a patchwork of sub-rules and processes that can readily be identified and networked into a coherent governance ‘paradigm’ for the technology. For example, as Marko Milanovic recently noted of AI in military contexts within the international criminal law paradigm:

[I]n my view, and when it comes to international criminal justice in particular, there is a tendency in some of the literature to overemphasize the extent of these challenges and the game-changing nature of AI as a disruptive new technology ...

[W]e’re going to be fine with the law as it stands today, at least insofar as the core business of international criminal courts and tribunals is concerned. AI does pose challenges, but they are not so radical and transformative that we won’t be able to effectively address them.⁴⁹

This does not mean that there is a single, clear, existing sub-rule for every current and anticipated issue—no rule set can achieve that. But the existing rules include criteria and mechanisms that also allow for any gaps or uncertainties in this patchwork paradigm to be resolved, and this is the second already existing component of a lifecycle governance regime for AI in military contexts without the need for new hard law.

Additionally, it is essential that what might be considered peripheral rule sets should also be mined for their utility in establishing governance frameworks. A case in point is contract law, which can operate both to constrain users (governments) and creators (enterprises) and to give effect to the ‘red lines’ of both the regulator and the regulated. In the first instance, contract is a sound method by which governments can build into development programs the limitations that international law imposes upon them. In the second instance, as the Pentagon–Anthropic dispute illustrates, enterprises can also seek to leverage contracts to place limitations on governments and militaries in terms of the uses that AI is put to.⁵⁰

One further benefit available from the existing rule sets is the extent to which parallel existing capabilities in interpreting and applying these rule sets—for example, military and government legal advisers and legally aware operators—are essential to this endeavour. The application of human experience and expertise in the initial acceptance (e.g., article 36 weapons review), and lifecycle monitoring as against the relevant governance framework will be essential to building trust in respect of AI in military contexts.

The Existence of Interpretive Mechanisms to Evolve Existing Law in the Face of New Technologies

International law already encompasses quite liberal systemic rules on interpreting extant law to fill gaps or clarify uncertainties. These range from the textual rules found in the VCLT,⁵¹ to the subsequent state practice rules,⁵² through to the rules for identifying customary international law (CIL)⁵³—noting that there is scope for the rapid formation of CIL where critical advances outrun the usual pace of CIL formation, as with Nuremberg,⁵⁴ space law,⁵⁵ and thus potentially AI law. That is, the existing systemic rules around international law interpretation are fundamentally evolutive rather than treated as frozen in time in an originalist manner. This allows for rules to be applied by analogy, by extension or by reduction from general principles in order to ensure coherent responses to new challenges, including the advent of new technologies requiring regulation in a manner that is coherent with existing paradigms.⁵⁶ The ‘how’ of this process is often a function of non-treaty negotiation fora such as soft law development, which is the third already existing component that supports the development of a lifecycle governance regime for AI in military contexts without the need for new hard law.

The General and Highly Useful Permissiveness Around Processes for Developing Soft Law

As noted above, soft law can be found in a wide variety of sources that go well beyond the more narrowly defined sources from which hard law can be distilled. Guidelines, declarations, resolutions, expert commentaries, statements of principle, codes of conduct, state non-papers and so on can all be used to develop the necessary interstitial threads required to bind together existing hard law and existing interpretive methodologies into a coherent regulatory scheme. And because the fora in which soft law instruments or processes can be developed are highly varied and permissive (which is not the case for hard law), using soft law as the regulatory glue for a lifecycle governance regime for AI in military contexts is achievable, acceptable and much simpler than seeking hard law responses to identified gaps or uncertainties. This is not to ignore the limitations that can attend soft law development—non-bindingness, like-mindedness, the risk of advocacy obscuring pragmatism—but it is inescapably the case that soft law (currently) offers a more realistic path to creating a lifecycle governance framework out of existing components than the negotiation of a new hard law instrument to that effect.

New Components Required

The previous section briefly described three existing components of the international legal system that can be—and are currently being—leveraged to develop a lifecycle regulatory scheme for AI in military contexts. However, this should not obscure the fact that there are a number of unknowns which will still need to be grappled to the ground so that the longevity of the lifecycle governance regime can be monitored and the next regime planned. That is, there is a point at which a lifecycle governance regime built on components we already have will fail because the ‘thing’ being governed will itself change so radically and qualitatively that it is no longer recognisable to (and thus capable of being regulated by) the applicable governance paradigm. This could perhaps be called, as *The Economist* noted of the advent of the Claude Mythos AI, a ‘Mythos moment’.⁵⁷ In this specific case, for example, it has been reported that the AI has not been publicly released precisely because its capabilities (including in respect of hacking and potential misuse in relation to biosecurity) are ‘unnerving’.⁵⁸ In this respect, the two key unknowns are the ‘what’ and the ‘when’.

Identifying the ‘What’ That Will Make an AI Regulatory Paradigm Shift Necessary

The first unknown is ‘the thing’—the ‘Mythos’ of the Mythos moment—that will change the paradigm as regards AI and military operations and capabilities. It is the advent or event that will render any existing or evolved governance approach qualitatively and systemically obsolete and ineffective. That is, we need to identify ‘the thing’—the incident, the initiative, the discovery, the accomplishment, the threshold—that will fundamentally change the nature, capabilities and/or latent malignity of AI from what we now see and what we realistically anticipate.

But identifying that thing is itself a problem. Is ‘the thing’ about the measure of trust in, or capacity to trust, an AI? Or is ‘the thing’ what happens when what goes on inside the black box is no longer explicable by or to humans? Or is ‘the thing’—as Sam White and others have argued—the point at which ‘apply[ing] anthropocentric law to a situation where damage is caused by this autonomous system’ becomes intolerably difficult or even impossible?⁵⁹ Or is ‘the thing’ something else entirely unforeseen at the moment? An example might be a globally condemned AI-executed strike that has disastrous consequences—a radically scaled-up version of the Minab school strike in Iran on 28 February 2026, which was linked to Claude and Maven.⁶⁰ A more tidal example might be the point at which an accumulation of reliance on AI by a military force manifests in a loss of receptiveness to political control, prompting governments to regulate more closely and intensely. Regardless of the nature of the ‘Mythos’ event, it is essential that someone, somewhere, is planning for the ‘what next’ once the Mythos—the paradigm-shifting ‘what’—occurs.

Identifying the ‘When’ of That ‘What’

Once the ‘what’ of the thing that will herald the necessity of a paradigm change is identified, we need to start thinking about the ‘when’. This is an identifiable point in time, technology and/or law where AI is approaching or achieves ‘the thing’—that is, the ‘moment’ aspect of the ‘Mythos moment’. This is because, at the culminating point of the anticipated ‘what’, any governance paradigm we evolve from the tools we currently have is likely to cease to be effective, or sufficiently effective, and to therefore need to be replaced. That is, the ‘when’ of the ‘what’ is the point in time of failure for the current regulatory paradigm. This may well be the point at which we ultimately conclude that we actually do need new hard law to underpin a (future) AI lifecycle governance regime, and therefore we ought to have some people already thinking—separately from the debate about the now—about this point in time.

Conclusion

It is a fact of the current geopolitical situation that the negotiation of new hard law to regulate AI in military contexts is almost certainly a non-starter. But it is also clearly arguable that (assuming political will) we already have the necessary tools to develop a functional and appropriate lifecycle governance regime for AI in military contexts. Further, it is likewise clear that in doing so, we ought to steer away from specific technology-based regulatory assumptions; rather, we should focus on processes, patchworks, soft law, existing and well-defined metrics, and what is anticipated rather than what is feared. However, there will inevitably come a point at which the technology that creates or supports the use of AI in military contexts will outstrip and render redundant any existing or derivative regulatory regime we can construct with the tools we already have. In order to prepare for that future, we should not let the (currently) enduring viability of any lifecycle governance regime we develop with those tools distract some of us from the need to also be looking at the ‘what’ and the ‘when’ of a future governance paradigm shift.

What might this mean for Army (and indeed for the ADF more generally)? Three thoughts are immediately apparent. First, a focus on creating a lifecycle governance framework out of existing tools, systems, rule sets, and architectures will require the development of deeply specialist legal officers who can understand and relate to AI operators in their own terms, and also AI operators who are legally aware and thus attuned to the governance potentialities—and threats—of both intra- and extra-paradigmatic ripples and developments. Second, wargaming the unlikely, the unexpected and the unwilling should be built into testing and analysis regimes so as to understand whether and when a paradigm-fracturing Mythos moment might eventuate—and thus be mitigated—regardless of the chance of actual occurrence. The lessons learned from wargaming

the worst scenarios from AI-linked operations will offer a sound basis from which to develop possible indicators and warnings that a Mythos moment may be on the horizon (and thus avoided). Third, further research into how use of AI in military contexts can be infused with chapeau influences such as command intent, political control and campaign appreciation should be prioritised. This is not an issue of tactical or meaningful human control; rather, it is an essential means to mitigate down to the lowest practicable level the possibility that the Mythos moment arises as a result of the military's operational loss of responsiveness to, or imbued-ness with, strategic and political oversight.

About the Author

Dr Robert McLaughlin researches, publishes, and teaches in the areas of Law of Armed Conflict, Law of the Sea, Maritime Security Law and Maritime Law Enforcement, and Military Law. He routinely engages in research activities, and course development and delivery, with the ICRC, the Australian Red Cross, the International Institute for Humanitarian Law, and the UN Office on Drugs and Crime. Rob came to the ANU College of Law after a career in the Royal Australian Navy as a Seaman officer and a Legal officer. Consequently, his research interests are primarily focused around issues of practical operational significance. His legal roles included as the Fleet Legal Officer, the Strategic Legal Adviser, as a Counsel Assisting the HMAS SYDNEY II Commission of Inquiry, Director Operations and International Law, and Director Naval Legal Service.

Connor McLaughlin is a legal officer and a research assistant at the Australian National Centre for Ocean Resources and Security at the University of Wollongong.

ENDNOTES

- 1 Natalie Nunn, 'How Lessons Learned from the CCW Can Help Guide the Regulation of AWS', *Lieber Institute Westpoint* (website), at: <https://lieber.westpoint.edu/how-lessons-learned-from-ccw-can-help-guide-regulation-of-aws>.
- 2 'Laws of War: Hague Declaration on the Use of Bullets Which Expand or Flatten Easily in the Human Body; July 29, 1899', *Yale Law School: The Avalon Project* (website), at: https://avalon.law.yale.edu/19th_century/dec99-03.asp; International Committee of the Red Cross (ICRC), 'Rule 77. Expanding Bullets', Customary IHL Study, *International Humanitarian Law Databases*, at: https://ihl-databases.icrc.org/en/customary-ihl/v1/rule77#refFn_BAA8276_00001.
- 3 ICRC 'Rule 77. Expanding Bullets'.
- 4 Convention on the Prohibition of Military or Any Hostile Use of Environmental Modification Techniques (10 December 1976), *International Humanitarian Law Databases*, at: <https://ihl-databases.icrc.org/en/ihl-treaties/enmod-1976>.
- 5 Silky Kaur, 'One Nuclear-Armed Poseidon Torpedo Could Decimate a Coastal City. Russia Wants 30 of Them', *Bulletin of the Atomic Scientists*, 14 June 2023, at: <https://thebulletin.org/2023/06/one-nuclear-armed-poseidon-torpedo-could-decimate-a-coastal-city-russia-wants-30-of-them>; Brandon Weichert, 'Russia's "Poseidon" Nuclear Drone Is Way Scarier Than You Think', *The National Interest*, 31 October 2025, at: <https://nationalinterest.org/blog/buzz/russias-poseidon-nuclear-drone-way-scarier-than-you-think-bw-103125>.
- 6 Emily Crawford, 'Accounting for the ENMOD Convention: Cold War Influences on the Origins and Development of the 1976 Convention on Environmental Modification Techniques', in Matthew Craven, Sundhya Pahuja and Gerry Simpson (eds), *International Law and the Cold War* (CUP, 2020), pp. 81, 94.
- 7 Callum Hamilton, comment on draft paper, 13 April 2026.
- 8 *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*, United Nations General Assembly Resolution A/79/239, para 7 (United Nations, 2024).
- 9 *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*, Report of the Secretary-General, A/80/78, para 57 (United Nations, 2025).
- 10 Australia, Canada, Czechia, Estonia, Japan, Latvia, Lithuania, Poland, the Republic of Korea, the United Kingdom and the United States, *Draft Articles on Autonomous Weapon Systems—Prohibitions and Other Regulatory Measures on the Basis of International Humanitarian Law*, CCW/GGE.1/2025/WP.7 (United Nations, 2025).
- 11 'Our Policy Position', *Stop Killer Robots* (website), at: www.stopkillerrobots.org/our-policies.
- 12 *Statute of the International Court of Justice*, article 38, no. 1.
- 13 Raluca Csernaton, 'Governing Military AI Amid a Geopolitical Minefield', *Carnegie Endowment for International Peace* (website), at: <https://carnegieendowment.org/research/2024/07/governing-military-ai-amid-a-geopolitical-minefield>.
- 14 'Artificial Intelligence in the Military Domain', *UN Office for Disarmament Affairs* (website), at: <https://disarmament.unoda.org/en/our-work/emerging-challenges/artificial-intelligence-military-domain>.
- 15 Daniel Thurer, 'Soft Law', in *Max Planck Encyclopedia of Public International Law* (2009).
- 16 Tan Hsien-Li, 'No Longer Hard Law's "Poor Relative": The Growing Respect for Soft, Non-Binding Legal Instruments in the International Order', *EJIL Talk!*, at: www.ejiltalk.org/no-longer-hard-laws-poor-relative-the-growing-respect-for-soft-non-binding-legal-instruments-in-the-international-order.
- 17 For example, see Mathew Montiel, 'Intangible Threats: How Uncontrolled Knowledge Fuels Proliferation', *Center for Arms Control and Non-Proliferation* (website), at: <https://armscontrolcenter.org/intangible-threats-how-uncontrolled-knowledge-fuels-proliferation>.

- 18 United Nations Arms Trade Treaty—for example, article 7(1)(b)(i) (export and export assessment) (New York, 2 April 2013).
- 19 United Nations Treaty on the Non-Proliferation of Nuclear Weapons (London, Moscow and Washington, 1 July 1968).
- 20 United Nations Comprehensive Nuclear-Test-Ban Treaty (New York, 10 September 1996).
- 21 Agreement Between Australia and the Agency for the Application of Safeguards in Connection with the Treaty on the Non-Proliferation of Nuclear Weapons (International Atomic Energy Agency, 10 July 1974).
- 22 United Nations International Convention for the Suppression of Acts of Nuclear Terrorism (New York, 13 April 2005).
- 23 Protocol of 2005 to the Convention for the Suppression of Unlawful Acts Against the Safety of Maritime Navigation (14 October 2005).
- 24 International Atomic Energy Agency, Vienna Convention on Civil Liability for Nuclear Damage (Vienna, 21 May 1963).
- 25 International Atomic Energy Agency, Convention on the Physical Protection of Nuclear Material and 2005 Protocol (Vienna, 26 October 1979).
- 26 International Atomic Energy Agency, Regulations for the Safe Transport of Radioactive Material (2018).
- 27 For example, the Comprehensive Nuclear-Test-Ban Treaty, UNGA Resolution 78/66, OP 4 (United Nations, 2023): ‘Urges all States not to carry out nuclear-weapon test explosions or any other nuclear explosions, to maintain their moratoriums in this regard and to refrain from acts that would defeat the object and purpose of the Treaty, while stressing that these measures do not have the same permanent and legally binding effect as the entry into force of the Treaty’.
- 28 For example, IAEA Code of Conduct on the Safety and Security of Radioactive Sources (International Atomic Energy Agency, 2004).
- 29 For example, Benjamin Jensen and John Paschkewitz, ‘Mosaic Warfare: Small and Scalable Are Beautiful’, *War on the Rocks*, 23 December 2019, at: <https://warontherocks.com/mosaic-warfare-small-and-scalable-are-beautiful>; Benjamin Jansen and Jake Kwon, ‘The US Army, Artificial Intelligence, and Mission Command’, *War on the Rocks*, 10 March 2025, at: <https://warontherocks.com/the-u-s-army-artificial-intelligence-and-mission-command>.
- 30 Australia, *Australia’s System of Control and Applications for Autonomous Weapons*, CCW/GGE.1/2019/WP.2/Rev.1 (Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 26 March 2019).
- 31 Ibid., para 3.
- 32 See Damian Copeland, Rain Liivoja and Lauren Sanders, ‘The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems’, *Journal of Conflict and Security Law* 28, no. 2 (2023): 285, 298–300.
- 33 *Defence Artificial Intelligence Strategy* (UK Ministry of Defence, 2022), para 5.1.1.
- 34 Alexander Blanchard, Vincent Boulanin, Laura Bruun and Netta Goussac, ‘Dilemmas in the Policy Debate on Autonomous Weapon Systems’, *SIPRI* (website), at: www.sipri.org/commentary/topical-back-grounder/2025/dilemmas-policy-debate-autonomous-weapon-systems.
- 35 Department of Defense, ‘DOD Adopts Ethical Principles for Artificial Intelligence’, media release, 24 February 2020, at: www.war.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence.
- 36 Shane Reeves, Ronald Alcalá and Amy McCarthy, ‘Challenges in Regulating Lethal Autonomous Weapons Under International Law’, *Southwestern Journal of International Law* 27, no. 1 (2020): 101, 117.
- 37 CA Ford, ‘AI, Human-Machine Interaction, and Autonomous Weapons: Thinking Carefully About Taking “Killer Robots” Seriously’, *New Paradigms Forum*, at: www.newparadigmsforum.com/p2526.
- 38 Christopher Menadue, ‘Science Fiction Helps Us Deal with Science Fact: A Lesson from Terminator’s Killer Robots’, *The Conversation*, 23 August 2017, at: <https://theconversation.com/science-fiction-helps-us-deal-with-science-fact-a-lesson-from-terminators-killer-robots-50249>.
- 39 Jean-Baptiste Jeangène Vilmer, ‘Terminator Ethics: Should We Ban “Killer Robots”?’, *Ethics and International Affairs*, 23 March 2015, at: www.ethicsandinternationalaffairs.org/online-exclusives/terminator-ethics-should-we-ban-killer-robots.

- 40 John Lewis, 'The Case for Regulating Fully Autonomous Weapons', *Yale Law Journal* 124 (2015): 1313, at: <https://yalelawjournal.org/comment/the-case-for-regulating-fully-autonomous-weapons>.
- 41 See, for example, 'Key Elements of Meaningful Human Control: Background Paper', *Article 36*, at: www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf; Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS), *Rolling Text* (GGE LAWS, 2025).
- 42 Vincent Boulanin, Neil Davison, Netta Goussac and Moa Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control* (SIPRI/ICRC, 2020), p. 3.
- 43 Gerald Mako, 'Legal Accountability for AI-Driven Autonomous Weapons', *Lieber Institute Westpoint* (website), 9 March 2026, at: <https://lieber.westpoint.edu/legal-accountability-ai-driven-autonomous-weapons>.
- 44 See, for example, James Kraska, 'Distinction, Proportionality, and Precautions in Attacks at Sea in the New Era of the Law of Naval Warfare', *Chicago Journal of International Law* 26, no. 1 (2025): 387, 391: 'The standard at sea is more lenient because vessels and aircraft, rather than individuals, are targeted. Belligerents are more likely to injure civilians during attacks on land than they are to injure civilians or civilian ships while attacking a warship or other military objectives at sea.'
- 45 European Union Artificial Intelligence Act, Regulation (EU) 2024/1689, Chapter II, article 5.
- 46 Robert Kolb, 'Indirect or Reverberating Excessive Collateral Damage in Modern IHL', *Lieber Institute West Point* (website), at: <https://lieber.westpoint.edu/indirect-reverberating-excessive-collateral-damage-modern-ihl>.
- 47 For example—albeit controversially—Nils Melzer, *Interpretive Guidance on the Notion of Direct Participation in Hostilities* (ICRC, 2009).
- 48 *Australia's Submission to the United Nations Secretary-General's Report on Lethal Autonomous Weapons Systems*, ODA-2024-00019/LAWS (Australian Government Department of Foreign Affairs and Trade, 2024), para 13.
- 49 Marko Milanovic, 'AI and the Commission and Facilitation of International Crimes: On Accountability Gaps and the Minab School Strike', *EJIL Talk!*, 9 March 2026, at: www.ejiltalk.org/ai-and-the-commission-and-facilitation-of-international-crimes-on-accountability-gaps-and-the-minab-school-strike.
- 50 'Statement from Dario Amodè on Our Discussions with the Department of War', media release, *Anthropic* (website), 26 February 2026, at: www.anthropic.com/news/statement-department-of-war.
- 51 Vienna Convention on the Law of Treaties (Vienna, 23 May 1969), articles 31–32.
- 52 For example, International Law Commission, *Draft Conclusions on Subsequent Agreements and Subsequent Practice in Relation to the Interpretation of Treaties* (United Nations, 2018), conclusion 5 (conduct as subsequent practice), conclusion 8 (interpretation of treaty terms as capable of evolving over time).
- 53 For example, International Law Commission, *Draft Conclusions on Identification of Customary International Law, with Commentaries* (United Nations, 2018), conclusion 2 (constituent elements).
- 54 Michael Scharf, 'Accelerated Formation of Customary International Law', *ILSA Journal of International and Comparative Law* 20, no. 2 (2014): 305, 330–335.
- 55 Fabio Tronchetti, 'Customary International Law and Space Law', in Mahulena Hofmann and PJ Blount (eds), *Elgar Concise Encyclopedia of Space Law* (Elgar, 2025), Chapter 13.
- 56 For example, via a lifecycle-based continuing appreciation of article 36 review obligations—see other essays in this collection.
- 57 'The Mythos Moment', *The Economist*, 18 April 2025.
- 58 *Ibid.*, p. 9.
- 59 Jacques Hartmann, Eva Jueptner, Santiago Matalonga, James Riordan and Samuel White, 'Artificial Intelligence, Autonomous Drones and Legal Uncertainties', *European Journal of Risk Regulation* 14, no. 1 (2023): 31, section III.
- 60 Kevin Baker, 'AI Got the Blame for the Iran School Bombing. The Truth is Far More Worrying', *The Guardian*, 26 March 2026, at: www.theguardian.com/news/2026/mar/26/ai-got-the-blame-for-the-iran-school-bombing-the-truth-is-far-more-worrying.

/ BUILT IN OR BOLTED ON: INDUSTRY'S ROLE IN DESIGNING HUMAN–MACHINE INTERACTION IN AI-ENABLED LAND WARFARE

Lena Trabucco

Introduction: Human Interaction in AI-Enabled Land Warfare

In early 2026, the dispute between Anthropic and the United States Department of Defense (US DoD) revealed a certain fragility within industry–government relationships regarding the development of military artificial intelligence (AI). Put simply, the dispute resulted in Anthropic refusing to authorise the use of its models for two specific use cases—fully autonomous weapon systems and mass domestic surveillance. The basis for the refusal was that these use cases ultimately violated the company's usage policies and its assessment of what current AI technology can safely and reliably do.¹ The US DoD argued that, from its perspective, it was not appropriate for Anthropic to predetermine the use cases for the Pentagon as long as the models were employed for 'any lawful use', and internal policies and regulations would guide US DoD use of, for example, fully autonomous weapons.² When Anthropic refused this sweeping access, the department designated Anthropic a supply-chain risk—a label conventionally reserved for foreign adversaries and never previously applied to a US company—and directed federal agencies to cease using its products.³ However, equally important was the signal to the wider defence-technology industry—grant unrestricted access or face potentially significant consequences.

This episode has important implications for the future of government–industry relationships across the globe. The US DoD signalled its intent to employ AI in contested use cases, or at least to preserve the flexibility to do so, and was willing to impose serious consequences on an American company that declined. Industry, in turn,

now has legitimate reasons for caution in partnering with the US DoD or other defence organisations on the development of such systems, particularly the firms best positioned to deliver reliable AI.

Understanding the contours of these relationships matters because industry has an irreplaceable role in developing reliable, responsible and lawful AI systems. Anthropic's original position rested on two concerns: that frontier AI is not yet reliable enough for fully autonomous weapons to function predictably, marking a red line for the company, and that there is a significant risk that commanders would delegate targeting decisions to machines. As Anthropic CEO Dario Amodei explained, the issue is that there is a need to preserve 'the right for military officers to make decisions about war themselves and not turn it over to machines completely'.⁴ Clearly, one of the core elements of this dispute is the appropriate role for human decision-making and the question of who gets to decide the boundaries of human-machine interaction. And to be sure, the issue and understandings of human control are valid and have been thoroughly analysed elsewhere.⁵ However, the unique role of industry in curating a decision space, and the role of human judgement within that space, remains underexplored. This article addresses this gap with a particular focus on land warfare.

AI is already influencing military operations not only through autonomous platforms but also through a growing range of decision-support systems that inform human decision-making.⁶ In contemporary and emerging land operations, AI is already, or is likely to be, employed in a range of use cases, such as intelligence analysis, target recognition, sensor fusion, logistics, mission planning, and battlefield decision support, among others.⁷ This level of AI integration points to a core concern for armies, which is less a question of autonomy and autonomous weapons, and more a question of how human commanders and soldiers will operationalise human judgement if AI increasingly shapes the information on which operational decisions are based.

For the Australian Army, this question is especially pressing.⁸ The Australian Army and the ADF as a whole have repeatedly recognised the importance of addressing human-machine interaction at the outset. Put simply, the human-control issues most pressing in a land environment differ, or at least have different dimensions, from those in maritime or air operations. Land warfare is marked by rapidly changing conditions, degraded communication environments, and greater uncertainty and complexity, particularly in urban environments. The introduction of AI systems—both decision-support systems and weapon systems—carries risks and opportunities for human control in land operations. Those risks and opportunities are explored elsewhere; this article focuses on the role industry plays in addressing the human-control issues specific to land operations.

The Australian Army has been at the forefront of addressing these questions. Numerous doctrinal and policy sources note the importance of incorporating AI while preserving space for human judgement. For example, the ADF Concept for Command and Control of the Future Force recognises that emerging technologies are reshaping the conduct of command and control and that Defence must adapt its concepts, education and capability development accordingly in order to ensure operator awareness and understanding.⁹ Additionally, the Army's Accelerated Warfare: Futures Statement for an Army in Motion frames adaptation to rapid technological and strategic change as a core challenge for the Army, emphasising the need to connect strategy, capability and innovation more effectively.¹⁰ Following these larger perspectives, the Army's Robotic and Autonomous Systems Strategy calls for robotic and autonomous systems (RAS) to 'present information in a useable and intuitive way to the soldier'.¹¹ Finally, the recent policy strategy for responsible use of AI in defence makes it clear that AI is not a niche technical issue but part of a broader transformation in how the Australian Army is commanded, coordinated and employed.¹² All of this guidance is necessary for building the policy and doctrinal foundations necessary for responsible AI integration and curating human–machine interaction, but there undoubtedly needs to be more specific guidance on design parameters for industry to develop Australian military AI. This article takes a step towards defining what that guidance could include.

The Australian Army and partner nations are right to highlight the acceleration of operations, and equally right to ask harder questions about the integration of AI into land operations specifically. To be sure, there are domain-specific considerations regarding the deployment of AI and the role of human judgement. The Army must prioritise a more human-centric focus as it equips and prepares soldiers, whereas other domains, such as air and sea, are more platform-centric. This is not to suggest that there are no important implications for human–machine interaction, but there are notable differences in these approaches.¹³ This distinction is also important because Army systems must function in environments where the assumptions underpinning AI models' development and training are likely at their weakest. Land operations are often conducted under unstable conditions, such as intermittent communications, a contested electromagnetic spectrum, degraded sensor inputs, incomplete data and high risks of adversarial deception.¹⁴ If AI systems are not developed, trained and tested for such conditions, the assumptions that industry has established in system design are unrealistic for these challenging conditions. This would result in AI systems becoming less reliable at critical moments when commanders are under pressure to act. Human interaction, therefore, cannot be understood merely as a matter of preserving the opportunity for a human to decide, but rather as the ability of commanders and operators to employ informed judgement and retain authority when systems are uncertain, partial or wrong. Industry plays a critical role in achieving this.

This article argues that preserving human control in AI-enabled systems depends on designing human–machine interaction so that informed judgement is preserved in contested operational conditions. In this context, the central challenge is not whether the human stays somewhere in the decision chain, or ‘in the loop’, but whether the human can appropriately judge, interpret, challenge and override the system’s recommendations or actions when required. Because much of this capability is designed and delivered outside the Army itself, the defence industry—in Australia and abroad—plays a decisive role in shaping whether human control is built into the system or merely bolted on through the operator’s interaction with it after deployment.

As discussed further below, this article offers five design mechanisms for preserving informed human judgement: interface design, presentation of uncertainty, override architecture, transparency of data inputs and outputs, and fallback states under degraded conditions. These are not peripheral technical details; nor are they considerations to be left to industry alone, because they are the ultimately practical mechanisms through which human judgement is either preserved or eliminated.¹⁵

The article proceeds in six parts. The first examines how AI is likely to influence decision-making in contemporary and future land operations. The second considers why degraded and contested environments make human–machine interaction especially difficult in Army operations. The third identifies the design mechanisms through which industry implements the preservation of informed human judgement in practice. The fourth proposes three practical design principles for AI-enabled systems intended for land warfare. The fifth examines implications for the Australian Army specifically. And the sixth concludes.

AI and Decision-Making in Contemporary Land Forces

One of the most transformative effects of AI in land warfare is its integration into the processes by which commanders and operators perceive, interpret and act on information. AI-enabled systems are increasingly employed for tasks such as processing intelligence, sensor fusion, route planning, predictive maintenance, autonomous logistics, operational planning, mission simulation, and support for targeting decisions, to name a few.¹⁶ In other words, AI will have a hugely influential role in land operations through shaping how humans make decisions compared to other applications with autonomous systems. This distinction matters because it shifts attention from questions of machine autonomy to the equally important question of how machine outputs enter command processes and shape human judgement.¹⁷

Experts similarly argue that the Army’s future relationship with machines is likely to evolve from the use of tools to various forms of human–machine teaming, with important implications for trust, command relationships and organisational adaptation.¹⁸

These arguments suggest that future Army operations will not merely add AI onto existing command arrangements but will involve a more fundamental reconfiguration of how information is produced, shared, interpreted and prosecuted.¹⁹

This has several implications for land operations. First, AI will compress parts of the decision cycle by processing large volumes of data at a pace beyond the capacity of teams of analysts.²⁰ That will certainly increase the speed of detection, prioritisation and recommendation outputs. Second, AI can aggregate data from multimodal sources and produce a synthetic operational picture that is likely to be difficult to construct in real time.²¹ Third, AI may increasingly recommend courses of action, or have other predictive functions, rather than simply presenting information—examples include flagging likely targets, suggesting routes, identifying anomalies, or predicting likely enemy movement patterns. This is an important distinction in identifying what AI offers that is ‘new’ in data analysis and processing. The significance is not simply in the collection and presentation of information but in the incorporation of that information to suggest tangible action and become a contributing element of the decision space. At that point, the issue is no longer one simply of efficiency but of how human judgement is shaped by machine-generated outputs.²²

For land forces, the significance of this problem is heightened by the distributed and time-sensitive nature of operations. Tactical commanders often act with incomplete information, limited opportunity for deliberation, and uneven communication with operational command.²³ AI-enabled decision support introduced into this environment may provide substantial benefits towards addressing those challenges, but it may simultaneously create new forms of fragility and reliance. For example, an apparently high-confidence recommendation may obscure the degree to which the system relied on incomplete or degraded inputs.²⁴ In another example, a common operating picture could conceal the system’s uncertainty rather than showcase it.²⁵ In another case, a prediction model could produce a seemingly reasonable output when, in fact, its assumptions are no longer applicable. In each case, the human is still present, but the quality of control depends on whether the system reveals or hides its own limits.²⁶

This suggests that the key control issue in Army AI is not necessarily the allocation of authority between humans and machines but the quality of the human–machine relationship through which decisions are made. A system that produces recommendations in a way that supports situational awareness, reveals uncertainty and requires critical and informed judgement will ultimately strengthen command. A system that presents outputs as authoritative and contestable only to a limited extent will weaken it, even if a human formally approves every decision.²⁷ Therefore, commanders do not only need to know why a system behaved as it did after an action has occurred; instead, they need to be able to determine whether its recommendation should be considered at all.

This issue becomes even clearer when considering the specific conditions of degraded and contested land warfare. It is one thing for an AI-enabled system to perform effectively when data are abundant, networks are stable, and sensor coverage and output are reliable. It is another for that same system to remain usable when communications are disrupted, data feeds are partial, or adversaries deliberately manipulate the information environment. The next section, therefore, examines the operational conditions that make human control in land warfare uniquely difficult, and why those conditions should be treated as the baseline for industry to design for.

The Reality of Land Operations: Degradation, Uncertainty, Complexity

As mentioned, a significant point of vulnerability for militaries incorporating AI is that systems are often designed, developed, trained and tested by industry, which has its own set of assumptions and interests.²⁸ But those assumptions do not necessarily reflect the realistic environmental conditions of land warfare. In conditions marked by intermittent communication or disrupted sensor feeds, for example, the problem of human interaction becomes more acute, as the reliability of machine-generated recommendations becomes more uncertain in critical moments with intense time pressure, and when they stand to benefit most from AI support.²⁹

Therefore, it is a pressing and immediate challenge for the human decision-makers to understand what the system is doing when the information environment is fragmentary or corrupted. In those instances, human judgement may be weakened or eliminated when a system presents conclusions without making visible the degraded quality of its inputs, the uncertainty of its outputs or the assumptions underpinning its recommendations. In that sense, degraded environments do not merely create technical problems for AI-enabled systems; rather, they expose whether human interaction and judgement have been designed as an operational function or preserved only as a formal decision point.³⁰

For AI-enabled systems, the challenges of land environments create at least three interrelated challenges. The first is data degradation. The quality of performance of AI systems depends on large volumes of timely, high-quality input data to generate reliable classifications, predictions or recommendations. In contested environments, however, data may be incomplete, delayed, contradictory or intentionally manipulated. Under these conditions, a system is likely to produce outputs that appear reasonable and actionable, even though the information supporting them has deteriorated significantly, thereby compromising their accuracy and operational utility. What is unclear is whether the operator can perceive that the system's picture quality has degraded.³¹

The second challenge is uncertainty in interaction. Human factors research has long demonstrated that operators are vulnerable to over-reliance on automated systems, particularly when decisions must be made quickly and when machine outputs carry an appearance of authority.³² Some experts note the distinction between misuse, disuse and abuse of automation to show that poor human–machine relationships arise not because humans are absent from the decision process but because systems are trusted beyond their appropriate scope or employed in conditions for which they were not designed.³³ In contested land operations, these dynamics carry particular weight. Accelerated pace and tempo compress the time available for deliberation, while an incomplete information picture increases the cognitive pressure to treat machine outputs as a substitute for judgement rather than a support to and enhancement of it. There is a risk that this will result in a gradual shift in the character of human control.

The third challenge concerns machine confidence in complex operational environments. While the previous challenge focuses on how humans relate to AI systems, the opposite is equally important—how AI systems behave when the operating environment becomes unreliable. AI is valued in military applications precisely because it can manage complexity—sorting, prioritising and recommending at machine speed. Yet in combat conditions, that apparent reduction in complexity may itself introduce operational risk. AI systems trained on historical patterns may continue to generate outdated outputs when the environment has shifted in ways those patterns do not capture. For example, a predictive model trained on an adversary's past behaviour may continue to produce recommendations based on that training even if doctrinal shifts or other vital signals indicate that the adversary is breaking from past patterns. This introduces new risks for commanders when introducing AI systems and highlights the importance of recognising the alignment of AI systems with the environment in which they are deployed.

Industry has a critical role in addressing these problems. If AI systems do not embed human control mechanisms for degraded environments, human control will be strongest when it is least needed and weakest when it is most needed. By contrast, if systems are designed on the assumptions that communications will be intermittent, data will be partial and adversaries will actively manipulate the information environment, human–machine interaction can be made to support—rather than substitute for—judgement under precisely those conditions.

For the Australian Army, this point has direct capability implications. Australia is unlikely to deploy AI-enabled land systems only in benign or highly connected environments. Any serious consideration of future Army capability must assume operations over wide distances, reliance on coalition networks, and exposure to electronic warfare, cyber disruption, and degraded battlespace awareness. Under those conditions, the value of AI will depend not only on the sophistication of the model but also on whether the system is designed to communicate uncertainty, reveal degraded inputs and inform human

judgement when the technical and operational picture is incomplete. Human control, in other words, is tested not when systems perform well in ideal conditions, but when they continue to function effectively even as the basis for trusting them is unclear.

Industry Mechanisms to Preserve Human–Machine Interaction

Industry is critical to tackling these challenges, and if human–machine interaction in AI-enabled land warfare is to be genuine rather than symbolic, it must be deliberately built into the system through thoughtful design. In practice, many of the systems through which the Army will encounter AI are developed not by military organisations themselves but by defence companies, software providers or other research partners. As a result, the practical operationalisation of human interaction is often determined before the system is fielded, with industry ultimately deciding the parameters of system architecture, the design of the interface, the way data are presented, the conditions under which users can intervene, and the mechanisms by which the system fails or reverts under degraded conditions. Therefore, questions of human interaction are a matter not only for commanders and operators but also for engineers and designers, and industry is central to resolving them.

As discussed above, the degraded conditions that characterise land operations pose significant challenges for AI system performance. Ultimately, systems developed for the Army are likely to operate precisely in those conditions where AI reliability is under the greatest pressure. Industry therefore plays a critical role not only by supplying AI-enabled capability but also by embedding specific mechanisms that preserve human judgement in such operationally realistic conditions. This article addresses five mechanisms in particular: interface design, presentation of uncertainty, fallback modes, override architecture, and transparency of data inputs and outputs. The discussion of each mechanism also includes a brief hypothetical example to illustrate the design features in concrete terms.

The first mechanism is interface design. Human control depends heavily on how the system structures and bounds the operator’s decision space. The interface is a critical tool because it does not merely display information—it shapes what the operator notices, what appears important, which courses of action seem available and how much cognitive effort is required to assess an AI recommendation. Consider a concrete example: a route-planning tool used by a tactical commander under significant time pressure. If the interface presents a single AI-generated route as the recommended option, the commander is effectively being asked to approve a conclusion rather than make an informed decision. She may have no visibility into whether the route was generated from current or outdated data, whether it includes the latest reported enemy positions, or how

the system weighted competing factors such as speed, cover and network connectivity. While the recommendation appears sound, the amounts of time and cognitive effort required to question it are high. This is not to suggest that the opposite is better—human factors research shows that the solution is not to display more information. Overloading the interface with information such as system assumptions, confidence scores and data provenance risks replacing one human factors problem with another—overwhelming commanders with details when clarity is the priority. Therefore, the design challenge for industry is one of calibration—presenting information in a way that supports informed judgement without sacrificing the speed and cognitive efficiency that make AI decision support valuable in the first place. Getting that calibration right for combat environments requires deliberate, operationally informed design choices for realistic conditions.

The second mechanism is the presentation of uncertainty. One of the greatest risks in AI-enabled military systems is that outputs appear authoritative even when the information supporting them has degraded. In land warfare, this risk is acute because operators work under time pressure and have limited capacity to validate machine-generated recommendations against other independent sources. To be clear, the problem is not that AI systems produce uncertain outputs, because uncertainty is a guaranteed feature in a complex operational environment. The problem is that uncertainty is invisible to the operator who receives and interprets those outputs. For example, consider a targeting recommendation generated by a sensor-fusion system operating in a complex environment. In this case, the operator is unaware that one of the system's data feeds has dropped entirely and been replaced by stored data from several hours earlier, and that a second sensor has been intermittent for the past 40 minutes. The recommendation the system produces looks the same as it would if all feeds were live and in real time, with the same format and apparent certainty. The operator may have no way of knowing that the picture underlying the recommendation has deteriorated, and—under time-critical pressure to act—she is unlikely to pause and investigate. This example points to an important risk: not of the system being wrong but of the operator having no reliable means of knowing when to be sceptical. The solution is not necessarily to attach a raw statistical confidence score to every output, as research shows this may not provide more clarity to the operator in the field. Industry is better served by designing systems so that uncertainty is communicated in forms that are operationally meaningful. It may seem counterintuitive, but a system will be more helpful if it displays uncertainty and clearly preserves the operator's capacity to calibrate trust, apply judgement, and intervene when confidence is no longer warranted.

The third mechanism is the design of fallback states.³⁴ As discussed, AI-enabled systems will not always have access to stable communications, complete data or uninterrupted connectivity. Fallback modes are the conditions a system reverts to when its primary capabilities are no longer available—a degraded state that may offer less analytical

capacity or fewer options but keeps the system functional and, importantly, keeps the operator informed about what has changed.³⁵ A system that continues to produce outputs without telling the operator that its underlying picture has deteriorated may leave the operator in a worse position than if the system had never been introduced at all, because she is now operating without the situational awareness she would have developed through other means. Consider another example. An AI-enabled targeting recommendation system is operating in an urban environment. Over the course of an operation, the system has built a detailed operational picture of a city—identifying new construction, structural changes to buildings, and other features of the urban landscape relevant to targeting decisions. However, the system experiences a malfunction and reverts to its fallback state, losing the operational picture it has collected and defaulting to its baseline map data. That baseline reflects the city as it was when the system was originally configured, before it learned anything during the operation, including about new construction and other structural changes. A poorly designed fallback makes none of this visible to the operator. The system would continue to generate targeting recommendations, and the commander would have no indication that the picture underlying those recommendations no longer reflected the city's current layout. Strike recommendations drawn from that degraded picture carry both operational and legal risks—such as hitting structures whose function or occupancy has changed, or missing features of the urban environment that have since become operationally significant. A commander cannot assess those risks if she does not know the picture has changed. A well-designed fallback, by contrast, alerts the operator immediately to the malfunction, clearly identifies which data has been lost, and flags that recommendations are now being generated from baseline data that does not reflect the real-time operational picture the system had previously accumulated.

The fourth mechanism is the design of the override architecture. It is common to claim that human control is preserved because a system includes a system override mechanism, especially in 'human-in-the-loop' models. However, in practice, having an override mechanism and having access to the information needed to determine its appropriate use are two different things. The mere existence of the 'stop' button does not, in fact, preserve human control. The more fundamental design question is not whether an operator can physically activate an override but whether she has sufficient information to know when to do so. Consider a targeting-support system that produces recommendations a commander finds tactically suspect. If the override mechanism operates in isolation from the system's other data sources, that is, if activating it simply stops the current recommendation without telling the commander why the system arrived at that recommendation, what data it was drawing on or how confident it was in its output, the commander is being asked to exercise judgement without the necessary information. She may or may not override correctly, but either way, the decision is less informed than it should be. A well-designed override architecture needs to be closely integrated with

the system's other transparency functions. For example, when an operator considers intervening, the system might include the information most relevant to that decision, such as what the recommendation was based on, where the data inputs are strong or weak, and, perhaps most significantly, what the operational consequences of an override are likely to be.

The fifth mechanism is transparency of data inputs and outputs. Transparency here means designing the system so that the operator has sufficient context around where a recommendation came from and what its limitations are, enabling an informed decision about how much weight to place on it. There are two dimensions of input/output transparency that are most critical here. The first concerns the data used to train and test the system. A system learns what it is taught, and if the conditions reflected in its training data do not match those in which it is deployed, its recommendations may be unreliable in ways that are not immediately apparent to the operator. For example, imagine a target-recognition system trained predominantly on land-based environments with disproportionately sunny conditions. When that system is deployed on a morning of dense fog, it is being asked to assess and recommend targets in conditions its training data did not adequately represent. If the operator has no way of knowing that the system has limited exposure to low-visibility environments, she cannot account for that limitation in her own judgement—and a recommendation that would warrant scrutiny in fog is accepted with the same confidence as one generated on a clear day. The second dimension concerns the data it is using in real time. In a separate example, imagine a system that estimates civilian presence in an urban area by collecting and analysing mobile phone signals. In a busy marketplace during a local holiday, the volume of signals may exceed the system's capacity to fully process them, leading the system to underestimate the number of civilians present and to produce targeting information that does not reflect the actual civilian risk on the ground. If the operator is not told that the system's data collection was incomplete and that the picture of civilian presence it is working from is only partial, then she has no basis for treating that recommendation with the caution it warrants. This is not to suggest that a soldier in the field needs to understand how a model was built. She does need to know, in plain terms, whether the system was designed for the environment she is operating in and whether the picture it is working from reflects the situation as it currently stands.

Taken together, these mechanisms illustrate the critical role of industry in curating human–machine interaction and maintaining informed judgement. Interface design shapes which information is used and how decisions are framed; the presentation of uncertainty shapes how much confidence the operator should have; fallback states shape how systems remain usable and appropriate under degradation; override architecture shapes the decision space for terminating the system appropriately; and transparency of data inputs and outputs shapes whether operators can understand the basis and limits

of machine recommendations. Through these mechanisms, industry determines whether informed judgement is embedded in the capability itself or simply understood as the requirement for a human to ‘press a button’. But, as is made clear, it is not this simple.

Designing for Human Control: Three Design Principles for Land-Based AI Systems

The previous section illustrates the critical role of industry in curating human–machine interaction through particular design mechanisms. But industry is currently left to its own devices to make many of the design choices that will ultimately impact the manner of human–machine interaction and how well informed human judgement will be in practice. These outcomes will not simply emerge automatically through technological sophistication or cutting-edge innovation. Whether they occur depends on whether governments, defence organisations, industry and other stakeholders align their efforts early enough to shape systems to remain appropriate and reliable under operational stress. Building on the mechanisms identified in the previous section, this section draws out three principles that should guide industry development. The first is to provide clear top-down guidance early enough to shape design. The second is to re-evaluate industry assumptions in development. The third is to design data presentation to surface limitations.

Provide Clear Top-Down Guidance Early Enough to Shape Design

The first principle is that governments and defence organisations must provide earlier and clearer guidance regarding human–AI interaction in AI-enabled military systems, for both decision-support tools and weapon systems. Without such guidance, industry is left to interpret vague political language or fragmentary policy signals on its own. In that absence, companies will design against whatever procurement requirements already exist—which are likely to be outdated, inapplicable to AI-enabled decision-support systems or silent on questions of human control altogether. A particular risk is that without such guidance, companies will interpret human–machine interaction differently and the range of interpretations will be wide, with varying levels of usefulness. The systems that the Army acquires could reflect that variation, and the practical meaning of human interaction and the impact on human judgement will be shaped less by deliberate policy than by the accumulated design choices of developers working without adequate direction.³⁶

To be sure, top-down guidance does not need to constrain industry creativity. Rather, this approach allows industry to channel it appropriately. Clear guidance on expectations or operational requirements for human interaction does not narrow the design space in ways that would stifle innovation. On the contrary, it gives developers

a well-defined problem to solve, and difficult, well-defined problems are exactly the conditions under which engineering creativity tends to produce its most useful results. The five design mechanisms offer many such concrete, well-defined problems: how to communicate uncertainty in operationally meaningful terms, how to design fallback states that preserve operator awareness, how to build override architecture that works under dynamic battlefield conditions. These are genuinely hard challenges, and there is no single correct answer. Top-down guidance that frames human interaction as a concrete design requirement invites industry to utilise their expertise to solve these problems, rather than leaving human interaction as an afterthought addressed through compliance language at the end of a procurement process.³⁷

If defence organisations want industry to build systems that prioritise human interaction and support informed human judgement, they must signal that expectation early enough to influence design, architecture, interface and testing decisions. Existing policy instruments in the US and UK illustrate both the potential for defence organisations to take this approach and the current limits on how they can do so. For example, the 2023 US DoD Directive 3000.09 establishes senior-level review requirements for autonomous and semi-autonomous weapons but stops short of providing industry-level guidance to help operationalise ‘appropriate levels of human judgement’.³⁸ Similarly, the UK Ministry of Defence’s 2022 ‘Ambitious, Safe, Responsible’ policy sets out AI ethics principles but leaves implementation largely to developers.³⁹ For the Australian Department of Defence, the absence of equivalent guidance on AI-enabled system design represents an opportunity rather than simply a gap. The ADF has already recognised this challenge. The Army RAS Strategy calls for the Army to ‘understand the design parameters of AI technology and algorithms [that may be] embedded in the fielded capability [and] may require Australian-derived verifiable-by-design AI, rather than a “black-box” solution’.⁴⁰ Offering this top-down guidance would give Australia the chance to put many policies into practice and establish a foundation for the Australian defence industry to develop these capabilities aligned with Australia’s prioritisation of informed human decision-makers.

Re-Examine Industry Assumptions in Design and Development

The second principle is that industry must re-examine the assumptions underpinning the design, development, training and testing of AI-enabled systems intended for land operations. As discussed above, these systems are typically designed, developed, trained and tested by industry actors whose assumptions do not necessarily reflect the realistic environmental conditions of land warfare. For the Army, the relevant operational environment is not one marked by stable and reliable conditions. Hence, systems designed against assumptions suited to permissive or idealised conditions may perform well in demonstrations but prove unsuitable for real combat environments.⁴¹

Degradation is the clearest example of this problem, but it is not the only one. A system built on the assumption of stable connectivity and reliable sensor coverage will struggle with disrupted communications or a contested information environment. The same failure mode occurs in any context when a system's training or design assumptions are no longer valid. In such cases, the system does not necessarily fail outright, as it continues to produce outputs, but those outputs rest on assumptions that no longer reflect the operational environment. Re-examining assumptions, therefore, changes the standard against which capability should be judged. This requires industry to consider not only whether the system performs well under the conditions for which it was designed and trained but also whether its output remains appropriate and intelligible when those conditions no longer hold. A system that performs well against its own design assumptions but becomes overconfident or difficult to interpret once those assumptions break down does not preserve human judgement and instead poses substantial risks to the forces that use it. Therefore, industry's assumptions should be treated not as fixed premises to be addressed only if they fail in the field, but as standing questions to be revisited against the conditions the system will encounter. For the Australian Army, which must plan for operations across wide distances, degraded communications, electronic warfare and coalition network dependencies, few of the assumptions that hold in controlled simulations can be taken for granted in practice.

Design Data Presentation to Surface Limitations

The third principle is that data presentation should be designed to disclose the limitations of the information underlying a system's output, not simply of the output itself. When such systems are introduced, the significant risk is not necessarily machine error or failure; it is misplaced or unwarranted confidence. A system that produces authoritative-looking outputs while concealing the fragility of the information underpinning them necessarily undermines the operator's capacity for informed judgement. Designing to surface limitations means that systems should show the operator not only what the machine recommends but where that recommendation came from, how confident the system is, and why it has that level of confidence, in terms the operator can act on. This is essential in land warfare, where decisions are often made quickly with incomplete information, and the consequences of acting on a flawed picture can be severe.

How a limitation is communicated matters as much as whether it is communicated at all. Take, for example, the issue of uncertainty. A raw statistical confidence score—a percentage figure attached to an output—may not carry much operational meaning for an operator, as a percentage may not have the same weight for each individual.⁴² It may be more useful to use visual representations of uncertainty that are immediately intelligible in operational terms—ones that convey, at a glance, the basis and quality of the

recommendation rather than its statistical properties. The design of that representation is one of the most consequential choices industry makes in determining whether human judgement is genuinely informed.⁴³

This principle also has a direct legal dimension. The obligations that international humanitarian law (IHL) places on commanders—to distinguish targets, to assess proportionality and to take feasible precautions before an attack—all depend on the commander's having an accurate understanding of available information.⁴⁴ A system that presents a targeting recommendation with apparent precision while concealing that its underlying picture is based on incomplete, degraded or outdated data actively impedes the commander's ability to fulfil her legal obligations. Surfacing the system's limitations is therefore not simply a design preference or a human factors consideration; it is an important component of IHL compliance and operational effectiveness. A system that is technically capable of surfacing limitations but is not designed to do so intelligibly will not result in informed human judgement.

Each of these principles is deliberately broader than any technical feature set. They are intended to guide how Defence proactively communicates the kinds of systems it needs, how industry operationalises those needs, and how Army assesses whether AI-enabled systems genuinely support human judgement in land operations. Ultimately, human interaction cannot be preserved simply by keeping a human in the loop; it must be deliberately designed into the institutional setting, development assumptions and operational behaviour of the system itself. For the Australian Army, that is the more demanding standard but also the more effective one.

Implications for the Australian Army

The foregoing analysis carries three implications for the Australian Army and, more broadly, for the Australian Department of Defence as it implements its Policy Settings for Responsible Use of Artificial Intelligence in Defence.⁴⁵ Released in March 2026, the policy settings are Defence's first dedicated AI governance framework applying across the entire ADF and the wider Australian defence portfolio. They set out Defence's obligations for responsible AI use under three headings—lawfulness, adherence to values-based principles, and proportionate controls—spanning the full technology life cycle from research and design through to decommissioning. Each policy principle maps nicely to one of the design principles set out above and speaks to a different gap between Defence's stated commitments and the practical mechanisms through which those commitments must be realised in the systems Army acquires and operates.

The first implication concerns top-down guidance. The three principles from Defence's policy settings currently operate at a level of generality that leaves industry to interpret what they require in practice. These principles reflect the broader public debate about military AI, which continues to treat autonomous action—particularly autonomous weapon systems—as the principal threat to human control. However, for land forces, the more pressing risk is the inability to promote informed judgement with decision-support systems that never act autonomously, and generalised principles are poorly suited to identifying that risk before systems are fielded. Articulating requirements earlier—specifying not only that human interaction must be preserved but also what that means for practical design features such as interface design, fallback states and override architecture—would translate principle into specification and position Australia as a contributor to an emerging international conversation on responsible AI design and development.

The second implication concerns coalition operations. The Anthropic–DoD dispute offers a clear warning about how unevenly perspectives on human interaction may translate into standards that are likely to be distributed across allies and partners. When a major coalition partner is prepared to designate a domestic AI provider a supply-chain risk for declining to authorise certain use cases, the implications for interoperability could be significant. Coalition operations may increasingly involve systems built on different control assumptions, with different fallback behaviours, different override architectures, and different conventions for surfacing uncertainty or other limitations. For an Australian commander employing coalition decision-support tools, the question is not whether the partner system is simply reliable, but whether its design assumptions are intelligible and compatible with Australia's own legal and policy commitments.

The third implication concerns data presentation, which is closely linked to the principle of explainability set out in the Policy Settings for Responsible Use of Artificial Intelligence in Defence. That principle requires that the function and the relationship between inputs and outputs be traceable, and that systems meaningfully accommodate human–machine interaction. However, as this article has shown, this can only be achieved if industry designs traceability into the system from the outset rather than treating it as something to be demonstrated after the fact. Australia has already established the policy foundation necessary to move into the next iteration of guidance and to tackle design-level challenges, ensuring that Australian Army commanders and operators have the most reliable systems available for operations.

Conclusions

AI is often discussed in military terms as a question of autonomy: how much independence a system should have, whether humans remain in or on the loop, and where formal responsibility should lie when machines engage in autonomous behaviour. While these questions are undoubtedly significant, a more immediate challenge is how AI-enabled decision-support systems will preserve and inform human judgement under the degraded, contested and time-pressured conditions that characterise land warfare. As this article has argued, human interaction in that context is less about human presence than about whether the human–machine relationship has been deliberately designed to support judgement, understanding, and intervention when necessary.

The article advances three linked claims. First, industry has a critical role in preserving informed human judgement. Second, human–machine interaction is a critical mechanism of control, and mechanisms such as interface design, the presentation of uncertainty, fallback states, override architecture, and the transparency of inputs and outputs determine whether commanders can judge, question and redirect system behaviour in practice. Third, three design principles can guide government and industry in strengthening these capabilities. The Australian defence industry plays a central role, as these mechanisms are determined upstream through capability design rather than downstream through doctrine alone. The practical meaning of human control is therefore shaped not only by command protocols but also by the design decisions built into the systems the Army acquires.

If Australian Defence treats human interaction as a design requirement from the outset, it can shape a different model of military AI that supports informed command rather than displacing it, and in which human judgement remains strongest when operational conditions are at their worst. That is a more demanding approach to capability development, but it is also the one most consistent with the realities of land warfare, responsible AI and the needs of future Australian Army operations.

About the Author

Dr Lena Trabucco is an Assistant Professor at the Centre for Military Studies, University of Copenhagen, a Nonresidential Research Fellow at the Stockton Center for International Law at the US Naval War College, and a Research Fellow at the Tech, Law, and Security program at American University, Washington College of Law. Her research focuses on the governance of emerging military technologies, with particular attention to artificial intelligence, autonomous weapon systems, human control, and responsible innovation in defense. She holds a PhD in international relations from Northwestern University and a PhD in law from the University of Copenhagen.

ENDNOTES

- 1 'Usage Policy', *Anthropic* (website), at: www.anthropic.com/legal/aup (accessed 15 June 2026).
- 2 United States Department of Defense, *Autonomy in Weapon Systems*, DoD Directive 3000.09 (DoD, 2023).
- 3 Hasan Ali Kanu, 'Hegseth Doubles Down on Anthropic's Security Risk Designation', *Politico*, 4 June 2026, at: www.politico.com/news/2026/06/04/hegseth-anthropic-designation-supply-chain-risk-00951183.
- 4 Ibid.
- 5 See Lena Trabucco, 'AI-Enabled Autonomous Weapons and Human Control: Part I: Human Control and Machine Learning Design and Development', *International Law Studies* 106 (2025): 533–578, at: <https://digital-commons.usnwc.edu/ils/vol106/iss1/17>; Anna-Katharina Ferl, 'Imagining Meaningful Human Control: Autonomous Weapons and the (De-)Legitimation of Future Warfare', *Global Society* 38, no. 1 (2024): 139–155, at: <https://doi.org/10.1080/13600826.2023.2233004>; Filippo Santoni de Sio and Jeroen van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account', *Frontiers in Robotics and AI* 5 (2018), at: <https://doi.org/10.3389/frobt.2018.00015>; Heather Roff and Richard Moyes, 'Meaningful Human Control, Artificial Intelligence and Autonomous Weapons', briefing paper prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons (2016), at: https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Roff_Moyes_Meaningful_Human_Control-with-cover-page-v2.pdf; Michael Horowitz and Paul Scharre, *Meaningful Human Control in Weapon Systems: A Primer* (Center for a New American Security, 2015), at: www.cnas.org/publications/reports/meaningful-human-control-in-weapon-systems-a-primer; 'Killer Robots: UK Government Policy on Fully Autonomous Weapons', *Article* 36, 19 April 2013, at: <https://article36.org/statements-updates/killer-robots-uk-government-policy-on-fully-autonomous-weapons-2>.
- 6 Anna Nadibaidze, Ingvild Bode and Qiaochu Zhang, *AI in Military Decision Support Systems: A Review of Developments and Debates* (Center for War Studies, University of Southern Denmark, 2024), at: www.autonorms.eu/wp-content/uploads/2024/11/AI-DSS-report-WEB.pdf; Paul Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence* (New York: W.W. Norton, 2023); James Johnson, 'Artificial Intelligence and Future Warfare: Implications for International Security', *Defense and Security Analysis* 35, no. 2 (2019): 147–169, at: <https://doi.org/10.1080/14751798.2019.1600800>.
- 7 Australian Army, *Robotic and Autonomous Systems Strategy v2.0* (Canberra: Commonwealth of Australia, 2022), at: <https://researchcentre.army.gov.au/rico/robotic-and-autonomous-systems/robotic-autonomous-systems-ras-strategy>; Alex Neads, Theo Farrell and David J Galbreath, 'Evolving Towards Military Innovation: AI and the Australian Army', *Journal of Strategic Studies* 47, no. 5 (2024): 669–698, at: <https://www.tandfonline.com/doi/full/10.1080/01402390.2023.2200588>.
- 8 Australian Defence Force, *ADF Concept for Command and Control of the Future Force*, version 1.0 (Canberra: Commonwealth of Australia, 2018), at: https://theforge.defence.gov.au/sites/default/files/adf_concept_for_command_and_control_of_the_future_force_v1_signed.pdf; Australian Defence Force, *Concept for Robotic and Autonomous Systems* (Canberra: Commonwealth of Australia, 2020), at: <https://defense.info/wp-content/uploads/2021/06/ADF-Concept-Robotics.pdf>; Benjamin Wood, 'Operational AI Integration and Governance in the Australian Army', *Australian Army Journal* 21, no. 1 (2025): 117–147, at: <https://doi.org/10.61451/1235807>.
- 9 Australian Defence Force, *ADF Concept for Command and Control of the Future Force*.
- 10 Australian Army, *Accelerated Warfare: Futures Statement for an Army in Motion* (Canberra: Commonwealth of Australia, 2018), at: <https://researchcentre.army.gov.au/library/other/accelerated-warfare>.
- 11 Australian Army, *Robotic and Autonomous Systems Strategy*.
- 12 Department of Defence, *Policy Settings for Responsible Use of Artificial Intelligence in Defence* (Canberra: Commonwealth of Australia, 2025), at: <https://defencescienceinstitute.com/news/dept-of-defence-policy-settings-for-responsible-use-of-ai-in-defence>.
- 13 The author thanks a reviewer for bringing this point to her attention.

- 14 Mick Ryan, *Human-Machine Teaming for Future Ground Forces* (Washington DC: Center for Strategic and Budgetary Assessments, 2018), at: <https://csbaonline.org/resource/human-machine-teaming-for-future-ground-forces>; Jack Watling and Noah Sylvia, *Competitive Electronic Warfare in Modern Land Operations* (London: Royal United Services Institute, 2023), at: www.rusi.org/explore-our-research/publications/occasional-papers/competitive-electronic-warfare-modern-land-operations.
- 15 See also Trabucco, 'AI-Enabled Autonomous Weapons and Human Control: Part I'.
- 16 Anthony King, 'Digital Targeting: Artificial Intelligence, Data, and Military Violence', *Journal of Global Security Studies* 9, no. 2 (2024), at: <https://doi.org/10.1093/jogss/ogae009>; Heiko Borchert, 'The Very Long Game of Defense AI Adoption: Introduction', in Heiko Borchert, Torben Schütz and Joseph Verbovsky (eds), *The Very Long Game: 25 Case Studies on the Global State of Defense AI* (Cham: Springer, 2024), pp. 1–38, at: <https://doi.org/10.1007/978-3-031-58649-1>; Andreas Graae, *Servers Before Tanks? Defence AI in Denmark*, DAIO Study 23/18 (Hamburg: Defense AI Observatory, 2023), at: https://defenseai.eu/wp-content/uploads/2023/11/daio_study2318_servers_before_tanks_andreas_graae.pdf.
- 17 HW Meerveld, RHA Lindelauf, EO Postma and M Postma, 'The Irresponsibility of Not Using AI in the Military', *Ethics and Information Technology* 25, no. 14 (2023), at: <https://link.springer.com/article/10.1007/s10676-023-09683-0>; Avi Goldfarb and Jon R Lindsay, 'Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War', *International Security* 46, no. 3 (2022): 7–50, at: https://doi.org/10.1162/isec_a_00425; Nadibaidze et al., *AI in Military Decision Support Systems*.
- 18 Alex Neads, David J Galbreath and Theo Farrell, *From Tools to Teammates: Human-Machine Teaming and the Future of Command and Control in the Australian Army*, Australian Army Occasional Paper No. 7 (Australian Army Research Centre, 2021), at: <https://researchcentre.army.gov.au/library/occasional-papers/tools-teammates>; Neads, Farrell and Galbreath, 'Evolving Towards Military Innovation'.
- 19 Neads, Galbreath and Farrell, *From Tools to Teammates*.
- 20 James Johnson, 'Automating the OODA Loop in the Age of Intelligent Machines: Reaffirming the Role of Humans in Command-and-Control Decision-Making in the Digital Age', *Defence Studies* 23, no. 1 (2023): 43–67, at: <https://doi.org/10.1080/14702436.2022.2102486>; Goldfarb and Lindsay, 'Prediction and Judgment'; Scharre, *Four Battlegrounds*.
- 21 Emelia Probasco, Helen Toner, Matthew Burtell and Tim GJ Rudneret, *AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks* (Washington DC: Center for Security and Emerging Technology, 2025), at: <https://cset.georgetown.edu/publication/ai-for-military-decision-making>.
- 22 Goldfarb and Lindsay, 'Prediction and Judgment'; Johnson, 'Automating the OODA Loop in the Age of Intelligent Machines'.
- 23 Lena Trabucco, 'AI-Enabled Autonomous Weapons and Human Control: Part II: Human Control and Military Commanders', *International Law Studies* 106 (2025): 579–615, at: <https://digital-commons.usnwc.edu/ils/vol106/iss1/18>.
- 24 Mica R Endsley, 'Supporting Human–AI Teams: Transparency, Explainability, and Situation Awareness', *Computers in Human Behavior* 140 (2023), at: <https://doi.org/10.1016/j.chb.2022.107574>.
- 25 Ibid.
- 26 Ibid.
- 27 Ibid.; Kenneth Payne, *I, Warbot: The Dawn of Artificially Intelligent Conflict* (London: Hurst Publishers, 2021); Ben Shneiderman, 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy', *International Journal of Human-Computer Interaction* 36, no. 6 (2020): 495–504, at: <https://doi.org/10.1080/010447318.2020.1741118>.
- 28 Lena Trabucco and Esben Salling Larsen, *Artificial Intelligence in Command and Control* (Djøf Publishing and Centre for Military Studies, 2025), at: [https://cms.polsci.ku.dk/publikationer/ai-i-kommando--og-kontrolsystemer/Artificial Intelligence in Command and Control download.pdf](https://cms.polsci.ku.dk/publikationer/ai-i-kommando--og-kontrolsystemer/Artificial%20Intelligence%20in%20Command%20and%20Control/download.pdf).
- 29 Mary L Cummings, 'Automation and Accountability in Decision Support System Interface Design', *Journal of Technology Studies* 32, no. 1 (2006): 23–31, at: <https://doi.org/10.21061/jots.v32i1.a.4>.
- 30 Endsley, 'Supporting Human–AI Teams'; National Academies (2021).
- 31 Endsley, 'Supporting Human–AI Teams'.

- 32 Raja Parasuraman and Victor Riley, 'Humans and Automation: Use, Misuse, Disuse, Abuse', *Human Factors* 39, no. 2 (1997), at: <https://doi.org/10.1518/001872097778543886>; Raja Parasuraman and Dietrich H Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration', *Human Factors* 52, no. 3 (2010), at: <https://doi.org/10.1177/0018720810376055>.
- 33 Parasuraman and Riley, 'Humans and Automation'.
- 34 Saleema Amershi et al., 'Guidelines for Human-AI Interaction', *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery, 2019), at: <https://dl.acm.org/doi/10.1145/3290605.3300233>.
- 35 Tamsyn Edwards and Paul U Lee, 'Designing Graceful Degradation into Complex Systems: The Interaction Between Causes of Degradation and the Association with Degradation Prevention and Recovery', *2018 Aviation Technology, Integration, and Operations Conference* (AIAA Aerospace Research Central, 2018), at: <https://doi.org/10.2514/6.2018-3510>; Michael A Gerber, Ronald Schroeter and Bonnie Ho, 'A Human Factors Perspective on How to Keep SAE Level 3 Conditional Automated Driving Safe', *Transportation Research Interdisciplinary Perspectives* 22 (2023), at: <https://doi.org/10.1016/j.trip.2023.100959>.
- 36 de Sio and van den Hoven, 'Meaningful Human Control over Autonomous Systems'; Steven Umbrello, 'Coupling Levels of Abstraction in Understanding Meaningful Human Control of Autonomous Weapons: A Two-Tiered Approach', *Ethics and Information Technology* 23 (2021): 455–464, at: <https://link.springer.com/article/10.1007/s10676-021-09588-w>.
- 37 de Sio and van den Hoven, 'Meaningful Human Control over Autonomous Systems'; Umbrello, 'Coupling Levels of Abstraction'.
- 38 Department of Defense, *Autonomy in Weapon Systems*, DoD Directive 3000.09.
- 39 Ministry of Defence (MoD), *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence* (MoD, 2022), at: www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence.
- 40 Australian Army, *Robotic and Autonomous Systems Strategy*.
- 41 Goldfarb and Lindsay, 'Prediction and Judgment'; ML Cummings, *Artificial Intelligence and the Future of Warfare* (Chatham House, 2017), at: www.chathamhouse.org/2017/01/artificial-intelligence-and-future-warfare.
- 42 Monica Tataschiere, Luke Strickland and Shayne Loft, 'Transparency Improves the Accuracy of Automation Use, but Automation Confidence Information Does Not', *Cognitive Research: Principles and Implications* 9, no. 67 (2024), at: <https://doi.org/10.1186/s41235-024-00599-x>; Na Du, Kevin Y Huang and X Jessie Yang, 'Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming', *Human Factors* 62, no. 6 (2020): 987–1001, at: <https://doi.org/10.1177/0018720819862916>.
- 43 Tataschiere et al., 'Transparency Improves the Accuracy of Automation Use'.
- 44 Ian Henderson and Kate Reece, 'Proportionality Under International Humanitarian Law: The "Reasonable Military Commander" Standard and Reverberating Effects', *Vanderbilt Journal of Transnational Law* 51, no. 3 (2021): 835–855, at: <https://scholarship.law.vanderbilt.edu/vjtl/vol51/iss3/12>; Emma J Breeze, 'Duty to Act on Knowledge: Precautions, Intelligence and the Protection of Civilians in Armed Conflict', *Journal of Conflict and Security Law* 29, no. 3 (2024): 311–329, at: <https://doi.org/10.1093/jcsl/krae015>; Trabucco, 'AI-Enabled Autonomous Weapons and Human Control: Part II'; Renato Wolf, 'Applying Precautions in Target Verification with AI Decision Support Systems', *Israel Law Review* 58, no. 2–3 (2025): 282–305, at: <https://doi.org/10.1017/S0021223725100071>; Jessica Dorsey, 'The Erosion of Human(e) Judgement in Targeting? Quantification Logics, AI-Enabled Decision Support Systems and Proportionality Assessments in IHL', *International Review of the Red Cross* 107, no. 930 (2025): 1041–1071, at: <https://doi.org/10.1017/S1816383125100969>.
- 45 Department of Defence, *Policy Settings*.

/ INDUSTRY DEVELOPMENT OF MILITARY AI: OPERATIONALISING INTERNATIONAL HUMANITARIAN LAW ACROSS THE LIFECYCLE

Taylor Kate Woodcock

Introduction

Military innovations today are driven by technology companies. While the military-industrial complex is by no means new, the advent of data-driven applications of artificial intelligence (AI)—such as machine learning (ML) and large language models (LLMs)—has brought these companies to the fore in unprecedented ways. While the wide array of contemporary research on AI acknowledges that algorithms ‘shape how we understand the world and they do work in and make the world through their execution as software, with profound consequences’,¹ it is crucial to also examine *who* shapes these technologies. For militaries integrating AI capabilities, this question has legal implications. The choices made by industry actors in developing these systems shape their technical function, how they can be used, and the military practices in which they are situated, with direct consequences for compliance with international humanitarian law (IHL). This spans a wide range of interconnected, multi-purpose systems that support various capabilities, from intelligence data fusion and analysis, to target production and civilian detection, operational decision support, and logistical processing.

The development of AI capabilities is a distributed process involving myriad industry and military actors in various roles across different stages of the lifecycle. While states may and increasingly do rely on industry for the development, evaluation and maintenance of military AI systems, the concentration of knowledge and technical expertise within the private sector creates a structural challenge for discharging IHL obligations.

This article develops this argument in two parts. The first explores the role of industry in developing AI capabilities and the implications this holds for states' fulfilment of obligations under IHL. It argues that, given the central role of industry to the development of military AI capabilities and the fact that choices by industry actors shape the technical function of these capabilities, how they can be used and the broader practices in which they are situated, states have a legal obligation to operationalise IHL in development practices throughout the lifecycle. However, this obligation risks being constrained by dependency on the knowledge and technical expertise concentrated within industry, requiring efforts to ensure independence and capacity to respect IHL. The second traces this argument concretely across selected lifecycle stages, demonstrating how development choices by industry actors can be a source of risk for IHL violations and constrain how states are able to fulfil their due diligence obligations. For military organisations navigating relationships with industry, the ability to conduct operationally sound and lawful operations is at stake, requiring critical attention to not only how AI capabilities are used but also how they are developed.

State Obligations to Operationalise IHL in Industry-Led Development of Military AI

This section explores the role of industry in developing AI capabilities and the implications this holds for states' fulfilment of obligations under IHL. It argues that, given the central role of industry to the development of military AI capabilities and the fact that choices by industry actors shape the technical function of these capabilities, how they can be used and the broader practices in which they are situated, states have a legal obligation to operationalise IHL in development practices throughout the lifecycle, which is constrained by dependency on the knowledge and technical expertise of industry. First, I outline how industry is central to the development of military AI capabilities, though this is heterogeneous and diffuse, with corporations of different sizes and activities contributing in diverse ways. Next, I demonstrate how AI capabilities are the product of development practices in which diverse private actors iteratively make choices that shape the technical function of systems, how they can be used and the broader practices in which they are situated, with implications for the lawfulness of military operations. As development choices by industry shape AI capabilities and military operations more broadly, I argue that states are legally obligated to operationalise IHL throughout development practices across the entire lifecycle. Finally, I highlight that states' dependency on industry knowledge and technical expertise risks constraining their ability to ensure IHL compliance throughout the lifecycle and that they must structure the relationship with industry to ensure the independence and capacity to operationalise IHL throughout development..

Industry and the Development of Military AI Capabilities

Today, industry has become central to the development of military AI capabilities, though its role is heterogenous and diffuse with corporations of different sizes and activities contributing in diverse ways. AI itself is a broad umbrella term capturing a range of technologies. The companies involved in the development of military AI are therefore not homogenous but vary in size and role.² The development of military AI involves not just large multinational corporations, but also small- and medium-sized enterprises, often funded by venture capital.³ This ranges from the behemoths Google and Microsoft, Palantir and OpenAI, to much smaller startups such as Anduril, Shield AI and Skydio.⁴ It also encompasses a number of different types of tech companies, including data companies, platform providers, model and software developers, and cloud services. Amid these various types of companies involved in the development of military AI, there are also different models of development. This includes the co-development of AI systems in collaboration with militaries, the production of products sold off the shelf, and the filtering-down and redevelopment of civilian products into the military domain.

The scale of state investment in military AI at present is significant. A 2024 Brookings Institute study found that US Defense spending for AI increased threefold in 2022–2023 with a 1,200 per cent increase in the value of contracts.⁵ Most recently, this year Palantir entered an enterprise agreement worth up to US\$10 billion with the US Army, which consolidates 75 contracts to provide ‘rapid access to cutting-edge data integration, analytics, and AI tools’.⁶ Anduril was likewise awarded a US\$99.9 million contract by the US Army for a ‘next generation command and control prototype’.⁷ This year, NATO also acquired Maven Smart System from Palantir.⁸ The growing prominence of LLMs with the promulgation of Open AI’s ChatGPT series has also garnered interest in defence, with the launching of Project Lima in the US to develop military applications of LLMs,⁹ and OpenAI also being awarded a US\$200 million contract with the US in 2025.¹⁰ These trends are repeated in the Australian context, where in January 2026 it was reported that the Australian Defence Force (ADF) is investing AU\$40 million in emerging technologies and AI through the Advanced Strategic Capabilities Accelerator (ASCA) Emerging and Disruptive Technologies program.¹¹ Fourteen contracts were awarded to both universities and private companies, including Cortisonic and Swordfish Computing. An ADF press release on these contracts stated that the investment is intended to ensure that it can ‘make decisions with speed and accuracy in an increasingly complex information environment’.¹² Additionally, it was reported that the Australian Department of Defence awarded Palantir a AU\$7.6 million contract in February 2026,¹³ though there are growing concerns around Australia’s investment in the company.¹⁴

The current landscape of military AI highlights the central role of industry in shaping emergent AI capabilities. Critical development choices are primarily made by diverse industry actors, at times in collaboration with defence but largely outside of military

institutions. It is therefore important to pay attention to how industry contributes to the development of military AI capabilities and whether this has consequences for international law, as examined in the following sections.

AI Capabilities as the Product of Development Choices

AI systems are the product of a multitude of cumulative development choices made primarily by industry actors across the lifecycle. In the context of practices of development distributed across the range of corporations involved in producing AI capabilities, these choices are made by a diverse actors, from engineers and teams of data scientists to product managers. These development choices throughout the lifecycle iteratively shape technical systems, how they can be used, and the broader practices into which they are situated. These choices can therefore contribute to shaping military decision-making and thus the lawfulness of operations.

As AI systems are the product of numerous development choices, they reflect and embed human decision-making. AI systems are often considered as objective and neutral due to their statistical and data-driven nature.¹⁵ This reflects ‘a technologically-inflected promise of mechanical neutrality’, which is often promoted by industry developing these systems.¹⁶ Through development choices, these systems embed a number of human subjectivities, including biases, values, interests and assumptions.¹⁷ For example, how data is labelled and prepared has a significant impact on the output of an AI system. The tendency to over-trust and uncritically rely on AI systems, referred to as automation bias,¹⁸ further masks that AI systems are the product of human development choices that inform the outcomes they produce. As such, the technical features of AI systems should not be considered separately from their development practices.¹⁹

As AI capabilities become increasingly integrated into military processes, development practices—and the inherent subjectivities embedded within them—can have a significant impact on how militaries conduct their operations. Indeed, these systems are being introduced in order to ‘revolutionise’ warfare.²⁰ Yet reliance on AI capabilities for situational awareness and operational decision-making has implications for how militaries conduct lawful operations. The risk of IHL violations flows not just from the technical features of these systems but also from the development choices that shape them, how military practitioners interact with them, and broader targeting processes as a whole. However, connecting outcomes in the battlespace with the many development choices made by heterogenous and diffuse industry actors is extremely complicated. As put by Seaver, this involves conceptualising AI not as ‘standalone little boxes, but massive, networked ones with hundreds of hands reaching into them, tweaking and tuning, swapping out parts and experimenting with new arrangements’, in which ‘the tendencies of an engineering team are as significant as the tendencies of a sorting algorithm’.²¹

Consequently, while concerns for control and human judgement characterise debates on military AI, ultimately there is no single human ‘in the loop’, but various decision points distributed across different actors at lifecycle stages. This raises an urgent challenge for militaries seeking to adopt AI capabilities, requiring them to consider how international law can be respected not just for a ‘final’ AI system or during use but also in the numerous choices made by industry developers throughout the AI lifecycle.

State Obligations to Operationalise IHL in Industry Development Practices

While international law obligations for military operations primarily relate to operational decisions, the impact of development choices made by industry on legal compliance means that consideration of IHL only at the point of final review or deployment is already too late. The risk of violations of IHL are a result of not only the technical features of AI systems and how they influence decision-making but also the development choices that shape these systems, how they can be used by military practitioners, and the broader military processes in which they are situated. As such, IHL must be operationalised throughout the development of AI capabilities, including decisions on how technical systems are designed, how capabilities are specified, and whether certain applications or functionalities should be developed at all. This section argues that it is a legal obligation for states to ensure IHL is operationalised throughout industry-driven development practices.

While states are the primary bearers of international law obligations,²² the role of private military AI developers nevertheless has direct implications for IHL compliance and the risk of violations. Core obligations owed under the core IHL targeting principles of distinction, precautions and proportionality may be undermined through the use of AI systems,²³ particularly regarding limited accuracy, reliability and transparency of AI systems leading to unintended military engagements and challenging civilian harm mitigation.²⁴ Concrete examples of risks for IHL compliance based on specific development choices made by industry across the lifecycle are traced in more detail in the next section.

The characteristics of AI systems suggest that the fulfillment of core IHL targeting principles must be considered already during the development of AI-enabled military capabilities. AI systems can shape decision-making by creating conditions—including of speed, complexity and opacity—that limit deliberation, frame information for decision-makers, and prompt uncritical reliance on AI outputs through automation biases.²⁵ These systems have limited performance in terms of accuracy (including errors and hallucinations) and reliability (particularly in complex and dynamic environments such as in conflict situations).²⁶ Combined, these characteristics can challenge the ability of military practitioners relying on these systems to take reasonable efforts to distinguish lawful targets from protected persons, to take constant care throughout operations and feasible precautions in attacks to mitigate civilian harm, including the verification of targets, and to assess proportionality.²⁷ Ensuring military practitioners have the

capacity to deliberate over these legal determinations and critically assess AI outputs must therefore already be considered when AI capabilities are developed, including in the design of technical systems, the organisation of targeting processes into which they are integrated, and the configuration of human–machine interactions. Fulfilling core IHL obligations cannot be operationalised through a final review of an AI system or throughout deployment alone. Development choices shaping the limits of accuracy and reliability, and human–machine interactions, including the conditions of operational decision-making, the framing of information, and biases in end users, are already fixed through development. To the extent that this limits the capacity of practitioners to make lawful targeting decisions, the moment for intervention has already passed. Respect for the principles of distinction, precautions and proportionality therefore necessitates that states ensure development choices made by industry actors are consistent with the requirements of IHL.

Together with these core IHL principles, the obligation for states to operationalise IHL throughout development practices is grounded in their positive duty to ‘ensure respect’ for IHL under common article 1 of the Geneva Conventions.²⁸ This entails a due diligence obligation to ensure compliance with IHL. Due diligence obligations in international law are obligations of conduct, requiring states to make reasonable efforts to fulfil legal duties, without necessarily requiring that they bring about a certain outcome.²⁹ IHL contains a number of due diligence obligations including the duty to ensure respect for IHL, which entails efforts to address violations that result from the conduct of third parties, such as private corporations.³⁰ Given the distributed development choices that ultimately produce AI systems, this due diligence obligation necessarily extends throughout the development process, not only at the approval of acquisition or during deployment.³¹

In conjunction with the core IHL principles and the duty to ensure respect for IHL under common article 1, the duty for states to legally review military capabilities also reinforces that the lawfulness of AI capabilities be assessed and secured throughout development. Article 36 of the First Additional Protocol (API) to the Geneva Conventions requires states parties to conduct legal review of weapons, means and methods of warfare.³² There is debate about whether this obligation captures AI technologies that are not directly connected to the use of force, such as AI-enabled systems to support targeting decisions.³³ Nevertheless, as IHL compliance requires operationalisation prior to deployment of AI capabilities, including for decision-support systems, this therefore suggests that a legal review is required *throughout* development, whether under article 36 or otherwise.

The legal obligation for states to ensure that IHL is operationalised in industry development choices for AI capabilities is reinforced by the framework of international human rights law (IHRL).³⁴ IHRL applies to military operations during armed conflict alongside IHL, and during targeting is read in light of IHL such that violations of the

principles of distinction, precautions and proportionality constitute a violation of the right to life.³⁵ Nevertheless, the development of AI systems driven by industry occurs primarily (although not exclusively) during peacetime, reinforcing the relevance and applicability of IHRL. In its General Comment 36, the United Nations Human Rights Council has highlighted that human rights holds relevance for a number of situations associated with the development of military AI, including ‘the deployment, use, sale or purchase of existing weapons and in the study, development, acquisition or adoption of weapons, and means or methods of warfare’, indicating that states must assess the impact of these activities on the right to life.³⁶ For instance, human rights impact assessments have been proposed as one mechanism for operationalising throughout the lifecycle, including during development.³⁷ IHRL has a particularly robust framework of positive obligations for states to protect human rights (in the military context, arguably including core IHL protections), including from business entities. This is reflected in the Business and Human Rights Framework, under which the United Nations Guiding Principles on Business and Human Rights (UNGPs) serve as authoritative guidance for preventing and addressing business-related human rights violations, including in conflict-affected areas.³⁸ Due diligence obligations under IHRL require two key conditions: the reasonable foreseeability of the (risk of) harm and the reasonable capacity to intervene.³⁹ While there may be some limitations to establishing jurisdiction over the actions of private companies, responsibilities for diligent prevention and protection of human rights can nevertheless require states to regulate these companies and prevent them from harming the human rights of people situated outside of their jurisdiction.⁴⁰ Read together, the frameworks of IHL and IHRL reinforce positive obligations on states to ensure respect for international law throughout the development of military AI systems by industry actors.

Accountability frameworks under international law serve as a basis to hold states responsible and engage industry when the development of AI capabilities contributes to violations of IHL. State responsibility provides a basis for holding states accountable for violations of international law attributed to them,⁴¹ including violations of IHL that result from the conduct of private developers acting as organs or agents of the state, as well as state failures to exercise due diligence in ensuring IHL compliance in private sector development.⁴² Regarding developers themselves, criminal liability has also been explored as a possible means of holding private developers responsible for international crimes committed with autonomous weapons and overcoming a responsibility gap.⁴³

Corporate responsibility offers a further framework to guide industry actors developing military AI with respect to IHRL. Next to state duties to protect human rights, the UNGPs also set out corporate responsibility to respect human rights, which is operationalised through human rights due diligence in business activities and relationships.⁴⁴ For industry actors developing military AI this entails identifying, preventing and mitigating human-rights related risks, which co-applied with IHL suggests addressing the risks of

violations of the principles of distinction, precautions and proportionality. Finally, regional instruments such as the EU AI Act also govern corporate conduct regarding AI and, despite excluding high-risk applications such as defence, remain relevant for companies developing dual-use AI systems for both civilian and military purposes.⁴⁵

State Dependency on Industry and Constraints for IHL Compliance

Technical knowledge and expertise in AI systems are concentrated primarily within the private sector, creating a dependency of defence on industry with significant implications for states' ability to ensure IHL compliance throughout the development of military AI capabilities. Industry provides not only AI technologies but also the expertise and specialised knowledge for their evaluation and assurance, use, and ongoing maintenance. This knowledge relates to frontier AI techniques, computer science and systems engineering, with industry retaining access to training data, system architecture and source code. This knowledge and expertise increasingly makes industry integral to military operations in which these systems are deployed. This is evident with the role of the private sector offering subscription-style services to AI-powered platforms and systems, as well as the increasing role for forward-deployed engineers.⁴⁶ Incentives to embed the role of industry are reinforced by current discourse framing national security around the urgent need for AI capability development in defence.⁴⁷

The centrality of industry for emergent military capabilities creates a distinct kind of procurement relationship. The concentration of knowledge and expertise on military AI systems within industry risks creating a dependency in which defence is reliant on the private sector.⁴⁸ This relationship of dependence can shape what kinds of AI systems are developed and procured by defence, as well as the standards, metrics and processes used to assess these systems. This can create challenges for defence in retaining adequate and independent oversight of AI systems, including for reviewing legal compliance. In particular, it may be difficult for defence to determine how development choices impact legal compliance without industry intervention.

Though states may outsource aspects of the development of military AI systems, they cannot outsource legal compliance and must still ensure respect for IHL during the development of these systems, as outlined in the previous section. Consequently, throughout the development of military AI systems, defence should maintain sufficient independence from industry, demanding transparency, preserving specialised knowledge to assess AI systems (either in-house or through independent institutions) and ensuring IHL is operationalised throughout all stages of the AI lifecycle. States retain critical knowledge regarding operational and legal requirements necessary for the development of military AI systems. As procurers, they can structure development practices with IHL obligations at the forefront by facilitating collaboration and knowledge-sharing with industry rather than a public–private relationship based on dependency. Industry should

likewise be concerned with IHL during development to provide operationally sound and lawful systems and avoid legal, commercial and reputational risk. Both defence and industry must therefore seriously consider and operationalise IHL throughout the development of military AI. To this end, these actors should be aware of how development choices throughout the AI lifecycle shape legal compliance. This paper serves to trace this by outlining how key development choices made at different stages of the lifecycle can impact compliance with the key IHL principles of distinction, precautions and proportionality.

Industry Development Practices Across the AI Lifecycle

Industry actors contribute to the development of military AI in a range of ways. Development practices are not isolated at one point pre-procurement but occur in a distributed and iterative fashion throughout the lifecycle of these capabilities. This section traces how industry development practices across the AI lifecycle can be a source of risk for IHL violations and constrain how states are able to fulfil their due diligence obligations. Building on engineering frameworks such as those developed in the IEEE ecosystem for the military AI lifecycle,⁴⁹ which identify lifecycle stages including planning, design, development, testing, deployment, operation and maintenance, this section situates these technical stages within a broader socio-technical lifecycle.⁵⁰ This spans the involvement of industry in the development of military AI capabilities spanning (1) governance, (2) research and development, (3) acquisition, procurement and review, and (4) deployment.

Governance

While it may seem counter-intuitive to suggest that the development of AI capabilities begins prior to research and development, in practice much of the ideation around AI capabilities begins already before any specific systems are developed. This occurs throughout the governance stage, where states and other stakeholders make efforts to govern, regulate and plan for the integration of military AI.

The military-industrial complex has been pervasive for some time, with lobbyists playing a more backward-facing role by meeting with states to promote industry agendas.⁵¹ Today, industry also plays a forward-facing role in some international governance contexts by actively participating in governance and the strategic planning of militaries on military AI, contributing to driving the conceptual space of warfare.⁵² Principally this takes place through the engagement of industry in global governance efforts and self-regulation initiatives. Industry contributes to the global agenda and norms around military AI by shaping understandings of what AI capabilities do, defining open concepts like ‘responsible AI’ and developing standards—including on legal compliance—before

states do so. In doing so, industry also contributes to shaping how compliance with IHL is understood and operationalised when decisions around AI capabilities are made during planning and governance efforts.

The need to involve industry among a range of relevant stakeholders in the governance of military AI has been recognised by both the United Nations⁵³ and other international efforts such as the Responsible AI in the Military Domain (REAIM) process.⁵⁴ Nevertheless, industry engagement at this stage is multivarious and uneven. As highlighted in a previous section, it involves a range of companies of different sizes and with different products, services and expertise. Various industry actors are present in multi-stakeholder state-led initiatives, such as the REAIM Summit series, and most recently with tech executives from OpenAI, Anthropic and Google attending the G7 Summit alongside state leaders in June 2026.⁵⁵ In contrast, there seems to be less industry engagement in some multilateral governance efforts, such as the United Nations Group of Governmental Experts on Lethal Autonomous Weapons (GGE LAWS) or the United Nations informal talks on military AI.⁵⁶

Industry also engages in global governance of military AI through self-regulation efforts. Industry self-regulation on AI involves the private sector developing and implementing its own rules, standards and best practices based on its specialised knowledge, expertise and practical experience.⁵⁷ This know-how is one of the main reasons why states engage with industry in governance debates on military AI.⁵⁸ Efforts at self-regulation may prove to fill a gap left by the governance stalemate in regulating military AI,⁵⁹ such as lack of consensus at the GGE LAWS,⁶⁰ and has been suggested to foster both innovation and ethics, legitimise developments and build public trust.⁶¹

The active participation of industry at governance fora on military AI and in efforts at self-regulation is exemplified at the series of REAIM summits hosted in The Hague in 2023, Seoul in 2024, and A Coruña in 2026. At each of these events, industry played a role both in the promotion of their products and in speaking on responsible innovation in the defence domain.⁶² For instance, in Seoul in 2024, the program featured 'REAIM Talks' to 'invite experts from both academia and industry to discuss how the policy discussion on responsible AI can be translated into engineering tasks'.⁶³ Demonstrating the intersection of military AI governance and industry self-regulation, in A Coruña in 2026⁶⁴ a representative of Microsoft took part in the plenary session on 'Framework for Responsible Industry Behaviour on AI in the Military Domain'.⁶⁵

Industry interventions in governance and self-regulation efforts can shape discourse and planning around military AI in line with corporate interests. After all, the models of innovation of companies developing military applications of AI are incentivised by national regulatory environments that are favourable.⁶⁶ It is in the interests of business development that industry props up discourse around AI as a necessary and urgent

requirement for national security, strategic acceleration and warfighting. The notion promulgated by the private sector that AI will win wars creates and legitimises a significant role for AI within military capabilities, and consequently of industry actors in their development. Schwarz highlights the point that by positioning AI developers as invaluable to future warfare over traditional arms manufacturers, industry both serves promises of long-term material benefits and establishes technical experts 'as experts in how to win wars'.⁶⁷

The pressures for rapid innovation cultivated by both states and industry risks sidelining efforts to understand the legal and ethical implications of these technologies.⁶⁸ This is not to say that industry obscures or ignores international law, but rather that by shaping discourse and setting standards, industry contributes to influencing understandings of legal obligations in a more subtle way. At times, industry has explicitly engaged with legal frameworks, such as IHL.⁶⁹ At the same time, initiatives such as the Framework for Responsible Industry Behaviour on AI in the Military Domain reflect joint efforts between international organisations, states and industry to develop a set of voluntary guidelines for the private sector that are not only grounded in existing legal frameworks but also notably 'emerging principles'.⁷⁰ In this sense, industry contributes to both interpretations of existing frameworks and the formulation of new norms on military AI, as reflected in the aim of the framework to 'clarify what we mean by responsible industry conduct ... grounded firmly in international law and norms'.⁷¹

This context highlights how the private sector is firmly embedded in the military AI agenda and is ostensibly indispensable to these efforts, giving it a platform through which corporate values such as efficiency can infiltrate discourse on the development and legality of these systems.⁷² The perceived need for rapid innovation mirrors and accentuates the need for efficient and expedited targeting processes. When AI capability development and targeting processes are increasingly optimised for efficacy and speed to better overcome the adversary, this embeds values that reflect 'logics of quantification' across the military AI lifecycle.⁷³ This quantification reflects the 'routinization' of military operations in which violence becomes procedural, emphasising efficiency, accuracy and speed.⁷⁴ These values around efficiency can further filter into how IHL is operationalised across the lifecycle.

At the governance stage, technical values driven by industry innovation and optimisation can shape how compliance with IHL is understood. IHL is a value-laden legal framework: while specific principles are often referred to as value laden, notably the principle of proportionality, this framework is also underpinned by the foundational principles of military necessity and humanitarianism.⁷⁵ These are values which IHL strives to balance. Yet when those with technical expertise within industry underscore the urgent need for AI capabilities, values of efficiency become instilled within what states define as 'militarily necessary'. Likewise with humanitarian considerations, the efficient streamlining of

targeting processes through AI capabilities is posed as a means of ensuring more precise targeting, protecting civilians and winning wars quickly. Yet there are serious questions as to whether the promise of AI lines up with its realities, given the inherent limitations of state-of-the-art AI systems in terms of accuracy, predictability and transparency.⁷⁶

Technical values around efficiency prioritised through corporate interventions in the military AI agenda can also contribute to compliance challenges for the IHL principles of distinction, precautions and proportionality. Based on industry knowledge and experience, weight and authority are given to industry claims about what AI capabilities do, even if untested. In some cases, AI systems are marketed as ‘battle tested’, for instance in the ongoing conflict in Ukraine,⁷⁷ implying that systems are operationally effective and can be used lawfully. However, whether AI systems can be used lawfully and have operational efficacy is highly context dependent and is not generalisable across different operational contexts.

Assumptions around AI integration from developers’ perspectives can misalign with legal requirements where developers view legal challenges as requiring technical fixes. For instance, while the development of AI capabilities can be based on assumptions that battlespaces can be adequately represented through data points and that decisions can be based on computational insights from that data, IHL requires contextual and value-laden decision-making that resists technical ‘fixes’ and requires consideration of the prevailing circumstances.⁷⁸ Another illustration is the development of AI systems using probabilistic risk scores to identify militants, as was reportedly done by Israel in Gaza.⁷⁹ Risk-based assessments of militancy do not account for the context-dependent assessment of whether a civilian is directly participating in hostilities or a member of a militant group is serving a continuous combat function.⁸⁰

Industry instilling the value of efficiency also influences ideas around how AI can alter the organisation processes of militaries and the professional roles of military practitioners within them, including for conducting legal assessments. Technical ‘fixes’ to make targeting more efficient can diminish the critical friction points of human deliberation over IHL, whether it be the status of a target, determination of feasible precautionary measures, or assessment of proportionality.⁸¹ These challenges present critical concerns around the ability of militaries to operationalise IHL when pursuing agendas of technological innovation and efficiency that are driven by industry. These shifts towards efficiency begin already during governance and planning efforts and are later reinforced by concrete design choices made during research and development, procurement and deployment.

Research and Development

Research and development are critical processes in which industry actors shape AI technologies, with consequences for compliance with IHL. Subjectivity and bias are entrenched in AI systems through their development, as discussed above in a previous section. Despite attempts to de-bias datasets, the subjective choices embedded in the development of ML models means that it is almost impossible to address and mitigate bias within these systems.⁸² For instance, Cummings and Li identify at least 12 choices during the construction of ML models that involve subjective choices by developers.⁸³ How AI systems shape behaviour and practices ‘inevitably reflect[s] the assumptions of the designer around who will use the system and what it should, will, and ought to be used for’.⁸⁴

Numerous development choices are made at the research and development stage that have the potential to shape how militaries comply with legal obligations and influence risks around violations. Sufficient quality and quantity of representative and context-appropriate datasets are critical for accurate performance of military applications of AI.⁸⁵ Choices of developers around what datasets to use and how they are labelled and prepared can influence how AI systems identify objects of interest and whether practitioners conduct targeting in compliance with the duty to distinguish lawful targets from protected persons and objects. Systems adapted from foundational models, such as adapting LLMs for targeting operations, can also raise challenges for IHL compliance, as the development choices made for a base model in the civilian sector may not be appropriate for the distinct environments, tasks and legal requirements in warfare. Furthermore, choices around specific mechanisms for explainability (such as more technical methods of xAI that provide computational explanations through probabilistically reverse engineering the system’s operation—or ‘model-of-a-model explainability’)⁸⁶ may also provide technical explanations that are not useful for the kinds of explanations around the basis for identification of a certain target that commanders need in order to substantiate their decisions and contest AI outputs.⁸⁷ Each of these examples highlights how choices made by developers can have cascading effects for IHL compliance once these systems are used. By way of illustration, the rest of this section examines development choices around training data to highlight why development choices are a source of risk for IHL violations, particularly around the duty of militaries to distinguish lawful targets from protected persons, principally including civilians.

The curation and preparation of training datasets are key stages where human choices inform how AI systems behave, with implications for compliance with IHL. For AI models trained through supervised learning approaches, the training data is determinative for the output of the system. Subjectivities and bias in the training data result not only from the data itself (a common concern surrounding the use of biased datasets) but also from the choices of those involved in preparing these datasets. Decisions around what data to

include, how to label, clean and encode it, and which relevant features to select reflect the ‘invisible labour’ that goes into the development of these systems.⁸⁸ The output of an ML model is therefore contingent on its training data, which is itself shaped by the processes in which it is labelled, selected and prepared to train and test the model.

Data labelling is a key part of this process, which occurs separately from the development of AI systems through a process in which data points are annotated with labels, often by companies operating under severe and exploitative working conditions.⁸⁹ There is a high degree of subjectivity in data annotation processes,⁹⁰ with data labelling taking place on the basis of natural language meaning. Much research points to the subjectivity during data annotation processes that leads to biased labels.⁹¹ The preparation of datasets by AI developers is a further process embedding human subjectivity in the development of these systems. There are various processes that developers go through to prepare datasets to train AI models. These include ‘data cleaning’, which involves subjective decisions around which data points should be included and excluded.⁹² They also include the process of feature selection, where data is selected based on what the developer deems to be relevant features. As the features selected are crucial for determining the eventual output of the system, feature selection is a significant site of human subjectivity in the development of AI systems.⁹³

Training data is critical to how AI systems behave. AI systems generate output based solely on the patterns identified from the training data, meaning that anything outside of this (referred to as ‘out of distribution data’) will not be accounted for.⁹⁴ The significance of training data for system output is demonstrated by the tank classifier parable, a story that explains how an artificial neural network classified pictures of tanks based not on the presence of a tank in an image, but on whether the image showed a sunny or overcast sky in the background.⁹⁵ There are several examples of this kind of ML error in reality, such as a classifier that distinguished huskies and wolves based on whether the image also showed snow.⁹⁶ These unintuitive errors highlight why the choices of developers around training data can significantly contribute to the outcomes of AI systems and how they perform in use compared to testing during development.

The significance of choices around training data has implications not only for system performance but also for risks around violations of IHL as a result. In particular, choices around the selection and preparation of training data can have implications for compliance with the IHL principle of distinction, requiring that militaries distinguish between lawful targets and protected persons and objects. Choices of developers on training data can contribute to unintended engagements and harm to protected persons when AI systems are relied on for targeting. These choices on data used to train AI systems contribute to determining the output the system will eventually produce and create the risks of introducing hidden assumptions, subjectivities and biases that skew AI output into the targeting process, leading to unintended engagements.

Decisions made according to the principle of distinction entail legal assessments that are distinct from development practices around data preparation.⁹⁷ For instance, determining whether individuals are combatants who are surrendering (*hors de combat*), civilians who are directly participating in hostilities, or militants who engaging in a continuous combat function requires contextual and qualitative legal assessments.⁹⁸ In contrast, data labelling is conducted not based on legal categories but from natural language understandings based on fragments of imagery interpreted outside of the context of armed conflict.⁹⁹ Likewise, data preparation is conducted by technical experts without domain expertise on military operations or IHL who are situated in a private civilian context rather than in armed conflict. To the extent that developers have preconceived ideas about what constitutes a combatant, militant or civilian or about which factors might be relevant for an AI system given its targeting function, choices around the labelling and preparation of training datasets can significantly contribute to the outcomes of AI systems. These hidden assumptions, subjectivities and biases can risk errors and unintended engagements resulting in IHL violations. Seemingly benign decisions about whether someone wearing a military uniform and bearing arms is a combatant, as well as biased judgements about what a militant ‘looks like’ based on race, gender and other discriminatory factors, reinforce biased ideas underpinning how targets are identified in armed conflict.¹⁰⁰

Whatever factors are deemed relevant by developers will inevitably shape the kinds of predictions an ML model will make. In data preparation practices like feature selection, the removal of factors that seem irrelevant to the developer may obscure the nuance and novelty that is characteristic of armed conflict and that must necessarily play into distinction assessments under IHL in targeting.¹⁰¹ Developers—who most likely will not have domain expertise on IHL—may, for instance, be unaware of the relevance of accounting for cultural idiosyncrasies, such as that in some contexts carrying weapons may be a typical cultural practice rather than indicating hostile intent and participation in hostilities. As a result, developers may discount the relevance of cultural factors reflected in certain data features; instead, they merely appear as features to be ‘cleaned up’.

The contextual nature of IHL distinction assessments highlights why the choice of appropriate training datasets for the context of use is crucial not just for operational accuracy but also for compliance with IHL. A challenge for developers is that in most cases it will be impossible for developers to know what training data will be appropriate for the context of use prior to deployment, yet this is critical for how the system will perform and whether it will result in false positives or false negatives.

Ultimately, whether unintended engagements connected to choices around training data for AI systems will constitute a violation of the IHL principle of distinction will turn on the reasonableness of the commander’s targeting decision. Given the tendency of AI systems to produce biased outcomes, in part due to the invisible choices of corporate actors

involved in data labelling and preparation, it is unlikely these decisions relying on AI output will be reasonable unless additional measures are taken to verify targets. However, challenges to verification exist when AI systems are used, notably due to the speed and scale of these systems and tendencies towards automation bias, as well as the difficulty of establishing the reliability of AI systems and predicting errors, particularly when these result from hidden development choices. The choices of developers around training data for AI systems therefore contributes to shaping outcomes in the battlespace and can be a source of risk of non-compliance with IHL.

Acquisition, Procurement and Review

During acquisition, procurement and review, industry plays a critical role in shaping how compliance with IHL is understood and assessed. Acquisition is a broad process in which states obtain military capabilities, generally encompassing the formulation of military requirements (identifying what capabilities are needed and prioritised), various models of development (by scientific institutions, state militaries and industry, increasingly in collaboration), and procurement processes (e.g. sourcing, tendering and contracting), as well as cross-cutting regimes for assurance and legal review. In light of accelerating innovation for military applications of AI, industry advocacy has become particularly influential in shaping and accelerating procurement processes.¹⁰² Along with the pressures around the urgent need for efficient integration of AI capabilities, in some instances this has resulted in the streamlining of acquisition processes.¹⁰³ While typically viewed as a separate lifecycle stage to the research and development stage addressed in the preceding section, in practice, new models of co-development between militaries and industry and streamlined procurement processes mean the boundary between these lifecycle stages is blurred. The development of AI systems by industry is linked to how militaries formulate their capability needs, what requirements they implement for tendering and contracting, how AI systems are tested and evaluated, and how legal reviews are conducted. Setting aside the highly relevant question of whether industry owes obligations during acquisition processes (e.g. through the business and human rights framework) for the purposes of this article, how states respect and implement IHL obligations is directly shaped by the role of industry. Nevertheless, states retain obligations to respect and ensure respect for international law, which cannot be outsourced to industry. This raises key questions about how IHL can be operationalised within acquisition processes that are structurally shaped by industry practices, expertise and interests.

The technical expertise of industry allows it a key role in structuring acquisition processes for military AI, with direct implications for how states fulfil obligations under IHL. Industry plays a determining role in shaping the specifications, metrics and methods of evaluation used to determine whether AI systems are fit for purpose. This also has flow-on effects

for legal reviews determining whether AI systems can be used in compliance with IHL. This section examines how industry's role in shaping acquisition processes for military AI, including technical assurance and legal review, creates structural risks for state compliance with IHL.

The development of frameworks for assessing AI systems is principally concentrated within industry and structures acquisition processes, including legal reviews. The assessment of AI systems is typically carried out through technical assurance processes, such as frameworks for testing, evaluation, verification and validation (TEVV).¹⁰⁴ From an engineering perspective, reliability testing involves determining the probability that a system will function correctly.¹⁰⁵ This requires a particular number of tests to gain statistical assurance of a required level of confidence in reliability.¹⁰⁶ Some have suggested that in order for the testing of military technologies to be meaningful, 'critical issues of performance must be translated into testable elements that can be objectively measured'.¹⁰⁷ However, the measurement of AI systems is never fully objective, involving assumptions and subjective choices of developers about how a system should function for specific tasks and how to measure this. Industry's technical expertise and access to proprietary systems (including source code), development processes and training data allow it to determine what specific metrics, methods and evaluation frameworks are used to assess AI systems.

AI raises a number of specific issues for testing, which are largely addressed through choices by industry actors. First, testing for accuracy and predictability is particularly difficult for contemporary AI systems as performance significantly degrades when the context of use deviates from the testing conditions or assumptions of developers.¹⁰⁸ This difficulty stems from the tendencies of AI systems to be 'brittle' (where small changes in input data compared to training and testing can significantly impact outputs)¹⁰⁹ and exhibit 'concept drift' (where real-world performance degrades as input data evolves over time),¹¹⁰ which are particularly challenging when AI systems are used in complex, dynamic or novel environments, all of which characterise conflict situations. Second, challenges for AI testing also relate to the limits of transparency in the design of these systems. The complex and opaque characteristics of AI systems that are unable to provide explanations presents challenges for understanding how the system functions and for testing as a result.¹¹¹ Third, difficulties also arise related to the large amount of data AI systems rely on, given that in the military domain pre-existing datasets and the opportunity for real-life testing are limited.¹¹² It may therefore require significant investment in methods for testing and evaluation, such as 'modeling, simulation, and experimentation', to ensure the system meets its specifications (verification) and functions as expected (validation), alongside rigorous testing and evaluation procedures.¹¹³ Testing methods that use synthetic data and simulations are choices made by developers that can contribute to compounding challenges around the gap between performance in testing and real-world use. Finally,

these challenges for testing AI capabilities are exacerbated by the disaggregated nature of AI systems and the broader networks and processes into which they are integrated. More specifically, the performance of different hardware and software components and the overall system must be assessed, along with the performance of the system in light of integration into broader ‘systems-of-systems’¹¹⁴ and organisational processes, including human–machine interactions. For instance, accuracy rates in testing will largely hinge on the source of input information, which may also be processed through other AI systems, and explanations will only offer traceability if meaningfully integrated into decision-making procedures and trusted by users.¹¹⁵

The selection of metrics and testing procedures for accuracy, methods for explanations to facilitate traceability in testing, and conditions for testing using synthetic data and simulations are all choices made by industry to enable them to test the systems they develop, and towards which developers aim for optimisation. These technical specifications are then relied on within acquisition processes.

Though technical assurance and legal review are distinct processes that serve different purposes, the technical assessment of AI systems nevertheless informs the review of their legality. As outlined above, technical assessment of AI systems during acquisition processes involves the technical expertise and choices of developers situated within industry. These technical assurance frameworks are developed to ensure systems are fit for purpose and operate as specified and expected. In contrast, legal review involves a determination on whether under some or all conditions military systems can be used in compliance with IHL, as well as other relevant frameworks such as IHRL. The review of weapons, means and methods of warfare under article 36 API must be undertaken at the stages of ‘study, development, acquisition or adoption of a new weapon, means or methods of warfare’, based on the anticipated ordinary use of the weapon at the moment of evaluation. This obligation applies equally to the states that research, develop, produce and sell weapons, as well as those that acquire them,¹¹⁶ with the latter having to conduct evaluations independently of any characterisation of legality purported to be held out by the exporting state or company.¹¹⁷

In order to review the legality of military AI capabilities, states must be able to assess how they function. Typically, one key aspect to ensure the review process is effective, among others, is to ensure ‘the availability of documentation on all aspects of the equipment to be procured which covers its military utility, its concept of use, capabilities and medical impact on the victim’.¹¹⁸ Such documentation includes the assessment frameworks developed by industry for assurance. The challenge is that technical assurance and legal review are distinct questions, which require different methodologies to address.

Crucially, legal obligations are not reducible to technical specifications. IHL obligations under the principles of distinction, precautions and proportionality that are held to the reasonable commander standard entail contextual and qualitative assessments.¹¹⁹ This includes whether a military objective serves a military advantage, whether civilians are directly participating in hostilities and whether combatants are surrendering or incapacitated (*hors de combat*), as well as how to take constant care to avoid civilian harm, take feasible precautionary measures in attacks, and avoid disproportionate attacks where expected incidental harm to civilians is excessive with respect to the anticipated military advantage. These discretionary assessments are held to a standard of reasonableness, and must be determined in the prevailing circumstances based on the information reasonably available. Legal review of AI systems therefore entails an assessment of whether in all or some conditions the use of these systems can be consistent with these obligations. This requires analysis beyond technical assurance.

Questions around what level of accuracy would be sufficient in order to be comply with the IHL principle of distinction, requiring that militaries distinguish legitimate military targets from protected persons and objects, are inherently limited. While accuracy is a useful technical metric under specific conditions, accuracy rates that are determined through the limited testing of AI systems in a particular context (often outside of conflict situations) do not map well to the highly contextual legal assessments needed to identify a military objective in armed conflict. Indeed, targeting according to IHL is not just a recognition task, and framing it as such betrays a reductive understanding of the contextual and qualitative assessments IHL requires.¹²⁰

The ability of a military commander to take feasible precautionary measures has the potential to be undermined by the conditions of speed, scale, complexity and opacity in targeting processes introduced by AI that result from concrete development choices. Respect for the principle of precautions arguably requires built-in friction points for deliberation over mitigating civilian harm in a manner that is sufficiently integrated into the organisation of command-and-control decision-making structures and addresses the challenges of uncritical reliance on AI through human–machine interaction.¹²¹ Moreover, the proportionality principle also entails a contextual and value-laden analysis that cannot be delegated to a computational system or traded as a proxy for AI-based collateral damage assessments.¹²² Legal review of AI systems must therefore account for how the use of AI systems shapes the agency of military practitioners and influences how they conduct the legal assessments they grapple with on a daily basis.

Ensuring respect for legal requirements requires states to operationalise IHL throughout acquisition processes. This includes considering the wide range of risks to IHL compliance, including bias, errors and hallucinations, probabilistic outputs, transparency, and impacts on speed and scale of targeting. Yet, industry shapes the assurance of AI systems in line with its own expertise and interests, prioritising technical requirements

over legal obligations. If states do not preserve the distinct aspects of legal review, this risks integrating an industry-driven, technically inflected approach to military capabilities at the expense of legal protections. While there may be challenges linked to lack of technical expertise within states conducting acquisition and lack of transparency within industry, states nevertheless have the obligation to ‘diligently seek information’.¹²³ States must therefore cultivate the capacity to understand the distinctions between technical assurance and legal review and preserve independence in assessing the legality of AI capabilities.

Because development and acquisition are deeply interconnected, this capacity for independent legal review cannot be reserved for the formal review stage alone.¹²⁴ As highlighted above, review of the legality of AI applications should occur throughout development processes, including at multiple stages during acquisition. Reviewing the legality of AI capabilities, including their integration into organisational targeting processes and resulting human–machine interactions, requires states to ensure that IHL considerations inform technical specifications and evaluation methods from the outset, rather than being applied only after industry has determined them. This requires the technical literacy not only to independently test industry claims about AI capabilities¹²⁵ but also to scrutinise, contest and redirect the methodological basis for such claims. Legal requirements should be a key component of understanding fitness for purpose: since the intended use of military capabilities should be lawful as stipulated in review requirements (e.g. under article 36 API), the specifications, methods and metrics used to assess AI systems cannot be determined using technical considerations alone. This also suggests that legal obligations can be a basis for the decision not to acquire a certain application or functionality. As highlighted by Goussac and Boulanin, interrogation of whether and why a certain military AI capability is needed is a critical aspect of responsible procurement.¹²⁶

Ensuring respect for IHL during acquisition of military AI necessitates a two-way knowledge exchange between states and industry on legal and technical expertise, and no review should be ‘compartmentalised’ on the basis of the different areas of expertise between lawyers, engineers, computer scientists and military stakeholders.¹²⁷ Flexible, collaborative and iterative acquisition and procurement processes have been suggested to address the complexities of military AI and address IHL compliance.¹²⁸ States must nevertheless lead this process without ceding authority to industry based on technical expertise and resources—what one commentator calls ‘building an intelligent customer capability’.¹²⁹ This requires states to structure the acquisition of military AI in a manner that centres respect for IHL. Given the complexity of negotiating assessments of technical performance and legal review, the question of precisely how states must operationalise these IHL obligations already during acquisition of AI systems is a critical area requiring further research.

Deployment

Development choices by industry post-deployment can shape how militaries operationalise IHL. A key characteristic of the evolution of military AI, and an area warranting much closer attention and research, is that the development of these capabilities is no longer fixed to a temporal stage prior to deployment but continues throughout the use in the battlespace. As with the development choices made through earlier stages of the lifecycle already discussed, when AI systems are being developed and used in the context of ongoing military operations, this can have critical implications for compliance with IHL.

In large part, development practices continue post-deployment due to the mutable character of these technologies. Software in general is subject to updates and iterative changes. The nature of contemporary AI systems is that they continue ‘learning’, in the sense of improving ‘performance after making observations about the world’.¹³⁰ For ML, ‘a computer observes some data, builds a model based on the data, and uses the model as both a hypothesis about the world and a piece of software that can solve problems’.¹³¹ To avoid performance degradation resulting from the brittleness and concept drift that AI systems are prone to (as mentioned above), models need to continue learning. Given the inherent risks of ‘online’ learning during the use of AI systems in warfare, systems are more likely to be maintained through ‘offline’ approaches, involving updates, retraining and fine-tuning. This necessitates development practices to continue post-deployment and builds in a role for industry actors. As highlighted above, emergent models of military–industry collaboration include the prominence of forward-deployed engineers and subscription-style services for AI systems.¹³² Throughout these processes, industry actors play an active role in redeveloping AI systems by engaging in maintenance activities and providing updates. This illustrates what Klonowska refers to as ‘tinkering’, which is a ‘highly experimental process of AI integration ... manifest[ing] through iterative re-training and fine-tuning of AI models, repeated changes to interface design, development, and adjustments of explainability features’.¹³³ This section examines two distinct illustrations of how industry development choices post-deployment can shape legal compliance, focusing on design decisions on how to address distribution shift in AI systems and corporate governance choices on what kinds of AI systems companies are willing to develop.

Distribution shift is a phenomenon that occurs when the scope of an AI system’s training data (the distribution) is no longer generalisable to the conditions of use, which can become significantly apparent after the system has already been deployed. As put by Wu, ‘the generalization performance of a learning algorithm is largely affected by distribution shifts between the training and test environments’.¹³⁴ This also impacts the real-world performance of AI systems where there is misalignment between training/testing data and the context of use. An example is if an AI system trained in one conflict environment is then deployed in a different conflict—for instance, in the snowy conditions of Ukraine compared to the different weather conditions in Iran.

AI systems will only reliably analyse situations based on patterns found within their training data, often generating unreliable or invalid output for situations that fall outside of that dataset. Distribution shift can raise particular challenges when general-purpose LLMs are used in the military domain but have been trained on large swathes of language-based data from the civilian domain,¹³⁵ requiring fine-tuning or other adaptation methods. A further difficulty is that distribution shift may be more difficult to detect due to the proneness of LLMs to confidently present invalid information (hallucinations) and the way these systems tend to appeal to the user (sycophancy).¹³⁶

If after deployment a model exhibits distribution shift, developers make choices about ways to (re)train, update or adapt systems for new data conditions, selecting from a number of available methods, such as transfer learning/domain adaptation, out-of-distribution generalisation, multitask learning, continual learning or meta-learning.¹³⁷ These are critical choices for the performance of the system and therefore how it impacts legal compliance.

Since the data distribution is a key determinant of how a system will perform in a given context, this has implications for assessments under IHL that are necessarily context dependent. When an AI system is updated to address distribution shift, the context in which the system is adapted (e.g. through retraining or fine-tuning on new data) directly informs legal assessments subsequently supported by the system. Distinction assessments on who counts as a combatant, militant or civilian directly participating in hostilities, and therefore a lawful target, or assessments on what feasible precautions are taken or if a proposed attack will be proportionate, all involve assessments based on the context at hand. How AI systems account for context can be shaped through the choices made by developers post-deployment, including how to address distribution shift. These approaches to the redevelopment of AI systems to address distribution shift shape how the information space is understood by commanders, highlighting (and obscuring) certain features of the environment and recommending particular courses of action over others, all of which shapes how IHL is operationalised during military operations. If military AI systems serve as a way to frame and filter information in a way that guides operational decision-making, then the choices made by both initial developers and those developing systems post-deployment determine the boundaries of that frame. When systems are the first point of inquiry—rather than legal advisors, for example—this positions those developing and retraining these systems as the experts, and highlights the significance of the decisions they make around how AI systems are updated. These decisions, such as the choice of data for retraining, can mean the difference between the identification of an individual as a militant or civilian, for example, shaping distinctions under IHL in targeting. There are also questions that arise when unintended engagements result from inaccurate information by AI systems as to the role of the developer who has been updating and adapting the system throughout deployment to address concerns about accuracy, and

how accountability is distributed and attributed to those responsible for any resulting violations of IHL, particularly for incremental development choices regarding learning paradigms that are hard to trace and understand.¹³⁸ Commanders and legal advisors must have the capacity to understand how updates to AI capabilities during deployment can alter their performance, the conditions under which they can be lawfully used, and the way they impact interactions with military practitioners. This also raises crucial questions about the need to continue conducting legal reviews of AI capabilities post-deployment.¹³⁹

A different kind of industry choice during deployment that impacts legal compliance is demonstrated by the 2026 controversy between the tech company Anthropic and the United States Department of Defense (DoD).¹⁴⁰ During the deployment of Anthropic systems, such as Claude, disagreement arose as to whether these systems could be used for certain purposes. Anthropic indicated that its systems could not be used for mass surveillance of the US population or as part of autonomous weapons systems, causing the DoD to cancel its contract with the company. This illustrates one way for industry to exercise agency by setting standards and self-regulating, even post-deployment. It demonstrates the fluidity between lifecycle stages for development choices made by industry that link to AI governance and self-regulation, procurement (through contract administration) and broader development choices about the kinds of systems they are not willing to produce. When a company like Anthropic sets its red lines, this may serve to legitimise other AI capabilities that are nevertheless problematic and raise legal challenges, such as AI-enabled decision-support systems. This article has highlighted many of the concerns around the development and use of AI capabilities, including decision-support systems, contributing to violations of IHL. Yet when industry sets its red lines only on domestic surveillance or autonomous weapons, these concerns are minimised. Moreover, other industry actors rapidly step in to fill the gap, as Open AI did in this situation, suggesting that understandings about the ethical desirability and legal compliance of certain applications will be interpreted differently by different industry actors, based on their commercial interests (whether reputational, ethical or commercial). This similarly occurred when Google employees protested at Project Maven being used for military purposes, leading to the company pulling out¹⁴¹ and eventually being replaced by Palantir, which developed Maven Smart Systems.¹⁴²

While they are significantly distinct processes, the illustrations of system updates for distribution shift and of self-regulation through choices not to develop certain applications show how industry development choices at different levels (one technical, the other corporate policy) and cross-cutting what are often treated as separate lifecycle stages (e.g. governance, procurement and development) shape how militaries comply with IHL. Development processes continue post-deployment, meaning that the choices industry actors make even after a system has been deployed can critically impact operational decision-making and the operationalisation of IHL therein. The values and interests that are embedded at the upstream stages of governance and planning, research and development, and acquisition are reinforced by these downstream development choices.

Conclusion

The development of military AI capabilities is not a neutral technical process. Industry actors make choices throughout the lifecycle of these systems that embed subjectivities, shape military operations, and have direct implications for respect for international law. Considering IHL compliance only in a final system review or during deployment is, therefore, already too late.

This article argues that states have a legal obligation to diligently operationalise IHL throughout the development of military AI capabilities, which cannot be outsourced to industry. While states may and increasingly do rely on industry for the development, evaluation and maintenance of military AI systems, the concentration of knowledge and technical expertise within the private sector creates a structural challenge for discharging IHL obligations. The article has demonstrated this by tracing how concrete development choices by industry across the lifecycle, from data collection and model design through to testing, integration and ongoing maintenance, can be a source of risk for IHL violations and can constrain how states are able to fulfil their due diligence obligations to ensure respect for IHL.

While existing debates on military AI and international law gesture to the need to consider IHL throughout the development of military AI, this article formulates this as a legal obligation and demonstrates concretely how development choices across different lifecycle stages shape the capacity of states to discharge it. This analysis is not intended to be comprehensive, but to demonstrate why industry development practices have direct implications for IHL compliance and that states are legally required to address these risks throughout the development of AI capabilities. There are myriad development choices across the lifecycle, as well as broader normative and epistemic questions about the agency of industry in the military domain, that warrant further scrutiny. Open questions remain, for instance, on whether industry actors developing military capabilities are directly participating in hostilities and owe legal obligations on that basis, or whether the reasonable commander standard may eventually need to account for a reasonable engineering standard.

For military organisations, including the Australian Army, the immediate implication is practical. Fulfilling legal obligations in relation to military AI requires cultivating the institutional capacity to scrutinise not only technical systems but also the development practices that shape system behaviour, how practitioners can use them, and broader operational processes. This means engaging with industry not from a position of dependency but with sufficient independence and specialised knowledge to direct development in accordance with legal obligations and operational requirements. As the development of military AI capabilities continues at an accelerated pace, this is a challenge that military organisations cannot afford to defer.

About the Author

Dr Taylor Woodcock is a researcher at the Asser Institute. Her research examines the implications of the development and use of military applications of artificial intelligence (AI) for international law, with a specific emphasis on International Humanitarian Law. She takes a practice-oriented approach to exploring the relation between international law and human agency, considering how the interaction between data-driven algorithms and military practitioners impacts this relationship. Her research was part of the research project *Designing International Law and Ethics into Military Artificial Intelligence (DILEMA)*, funded by NWO–MVI Programme on Responsible Innovation (2020–2024). Taylor has research experience in various fields of international law, including international humanitarian law, international human rights law, international criminal law, and international arms control and disarmament law. Taylor graduated cum laude from the Leiden University LL.M. specialising in Public International Law (2019). She also has a double degree in Law and International Studies (International Relations Major) as part of her undergraduate studies at Deakin University in Melbourne, Australia. Following her studies, Taylor completed a year-long internship at the Asser Institute in the research strand ‘Human Dignity and Human Security in International and European Law’. Her key areas of research in this position included international human rights law, international humanitarian law and international criminal law. She was also a pro-bono research assistant at the strategic human rights litigation practice, *Human Rights in Practice*.

ENDNOTES

- 1 Rob Kitchin, 'Thinking Critically About and Researching Algorithms', *Information, Communication & Society* 20 (2017): 14–29, 18.
- 2 Alexander Blanchard, Vincent Boulanin and Laura Brunn, 'Mapping the Military AI Industry', *Stockholm International Peace Research Institute* (SIPRI) (website), 23 April 2026, at: www.sipri.org/commentary/topical-background/2026/mapping-military-ai-industry.
- 3 Elke Schwarz, 'From Blitzkrieg to Blitzscaling: Assessing the Impact of Venture Capital Dynamics on Military Norms', *Finance and Society* 1 (2025): 1–24, 1.
- 4 Marijn Hoijtink and Jasper van der Kist, 'Platforms on the Frontline: The Rise of the Platform Model in Defense Tech', *Opinio Juris*, 11 February 2025.
- 5 Kevin C Desouza, 'The Evolution of Artificial Intelligence (AI) Spending by the U.S. Government', *Brookings Institute*, 26 March 2024, at: www.brookings.edu/articles/the-evolution-of-artificial-intelligence-ai-spending-by-the-u-s-government.
- 6 US Army Public Affairs, 'U.S. Army Awards Enterprise Service Agreement to Enhance Military Readiness and Drive Operational Efficiency', *U.S. Army* (website), 31 July 2025, at: www.army.mil/article/287506/u_s_army_awards_enterprise_service_agreement_to_enhance_military_readiness_and_drive_operational_efficiency.
- 7 'Anduril Awarded \$99.6M for U.S. Army Next Generation Command and Control Prototype', *Anduril Industries* (website), 18 June 2025, at: www.anduril.com/article/anduril-awarded-usd99-6m-for-u-s-army-next-generation-command-and-control-prototype.
- 8 'NATO Acquires AI-Enabled Warfighting System', *NATO* (website), 14 April 2025, at: www.ncia.nato.int/newsroom/news/nato-acquires-ai-enabled-warfighting-system.
- 9 US Chief Digital and Artificial Intelligence Officer (CDAO), *Task Force Lima Executive Summary* (CDAO, 2024), at: www.ai.mil/Portals/137/Documents/Resources%20Page/2024-12-TF%20Lima-Exec-Sum-TAB-A.pdf.
- 10 AFP, 'OpenAI Wins \$200m Contract with US Military for "Warfighting"', *The Guardian*, 17 June 2025, at: www.theguardian.com/technology/2025/jun/17/openai-military-contract-warfighting.
- 11 Christine Casimiro, 'Australia Plugs \$26M into AI to Supercharge Military Decisions', *Military AI*, 7 January 2026, at: <https://militaryai.ai/australia-defense-ai>.
- 12 Australian Government Department of Defence, 'Harnessing Tech to Drive Faster Decisions Across Defence', media release, *Defence* (website), 5 January 2026, at: www.defence.gov.au/news-events/releases/2026-01-05/harnessing-tech-drive-faster-decisions-across-defence.
- 13 Robert Dougherty, 'Palantir Secures \$7.6m Defence Contract to Supply ICT System Platform', *Defence Connect*, 17 February 2026, at: www.defenceconnect.com.au/industry/17697-palantir-secures-7-6-million-defence-contract-to-supply-ict-system-platform.
- 14 Josh Taylor, 'Calls Grow to Ban Palantir in Australia After Manifesto Described by UK MP as "Ramblings of a Supervillain"', *The Guardian*, 30 April 2026, at: www.theguardian.com/technology/2026/apr/30/palantir-manifesto-australia-government-contracts.
- 15 Neil Renic and Elke Schwarz, 'Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing', *Ethics & International Affairs* 37, no. 3 (2023): 321–343, 337.
- 16 Tarleton Gillespie, 'The Relevance of Algorithms', in Tarleton Gillespie, Pablo J Boczkowski and Kirsten A Foot (eds), *Media Technologies: Essays on Communication, Materiality, and Society* (Cambridge MA: MIT Press, 2014), pp. 180–181.
- 17 Mary L Cummings and Songpo Li, 'Subjectivity in the Creation of Machine Learning Models', *Journal of Data and Information Quality* 13, no. 2 (2021): 1–19; Gillespie, 'The Relevance of Algorithms', p. 178; Kitchin, 'Thinking Critically', p. 18.
- 18 Theo Araujo, Natali Helberger, Sanne Kruijemeier and Claes H de Vreese, 'In AI We Trust? Perceptions About Automated Decision-Making by Artificial Intelligence', *AI & Society* 35, no. 3 (2020): 611–623, 613.
- 19 R Stuart Greiger, 'Bots, Bespoke, Code and the Materiality of Software Platforms', *Information, Communication & Society* 17, no. 3 (2014): 342–356, 346.

- 20 See e.g. European Commission, *Artificial Intelligence (AI) in Defence* (European Union, 2025), at: <https://op.europa.eu/en/publication-detail/-/publication/a856c715-e075-11f0-8439-01aa75ed71a1/lan-guage-en>.
- 21 Nick Seaver, 'Knowing Algorithms', in Janet Vertesi and David Ribes (eds) *digitalSTS: A Field Guide for Science & Technology Studies* (De Gruyter Brill, 2019), pp. 412–422, 419.
- 22 Increasing attention has been paid to non-state actors as bearers of international law obligations: see e.g. James Summers and Alex Gough (eds), *Non-State Actors and International Obligations* (Leiden: Brill Nijhoff, 2018), at: <https://doi.org/10.1163/9789004340251>; as well as more specifically non-state armed groups and private contractors: see e.g. Daragh Murray, 'How International Humanitarian Law Treaties Bind Non-State Armed Groups' *Journal of Conflict & Security Law* 20, no. 1 (2015): 101–131; Emanuela-Chiara Gillard, 'Business Goes to War: Private Military/Security Companies and International Humanitarian Law', *International Review of the Red Cross* 88, no. 863 (2006): 525–572.
- 23 See Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, 1125 UNTS 3 (API), arts 48, 51, 57.
- 24 See e.g., Abhimanyu George Jain, 'Autonomous Weapon Systems, Errors and Breaches of International Humanitarian Law', *Journal of International Criminal Justice* 21, no. 5 (November 2023): 1005–1032, at: <https://doi.org/10.1093/jicj/mgad043>; Jessica Dorsey and Marta Bo, 'AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks and the Human(e) Dimension', *International Law Studies* 107 (2026): 137; Jessica Dorsey, 'The Erosion of Human(e) Judgement in Targeting? Quantification Logics, AI-Enabled Decision Support Systems and Proportionality Assessments in IHL', *International Review of the Red Cross* 107, no. 930 (2025): 1041–1071; Rainer Rehak and Taylor Kate Woodcock, 'Automating Civilian Harm: On Israel's Use of the AI-Enabled Targeting System Lavender in Gaza and International Humanitarian Law', in *FAccT '26: The 2026 ACM Conference on Fairness, Accountability, and Transparency* (New York NY: Association for Computing Machinery, 2026), pp. 4885–4898, at: <https://doi.org/10.1145/3805689.3812357>; Taylor Kate Woodcock, 'Human-Machine (Learning) Interactions: War and Law in the AI Era' (PhD dissertation, University of Amsterdam, March 2026); Klaudia Klonowska and Taylor Kate Woodcock, 'Rhetoric and Regulation: The (Limits of) Human/ AI Comparison in Legal Debates on Military AI' in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (Asser Press, 2026).
- 25 Woodcock, 'Human-Machine (Learning) Interactions'; Taylor Kate Woodcock, 'Machine Learning: AI, Algorithmic Arbitrariness and Legal Decisions', in Matilda Arvidsson and Kieran Tranter (eds), *AI, Law and Politics* (Routledge, forthcoming 2026).
- 26 Klonowska and Woodcock 'Rhetoric and Regulation'; Rehak and Woodcock, 'Automating Civilian Harm'; Henning Lahmann, 'Bad Algorithms and the Epistemic and Discursive Powers of Military AI', *SSRN*, 29 October 2025 (forthcoming in the *London Review of International Law*, 2026), at: <https://papers.ssrn.com/sol3/Delivery.cfm/6392939.pdf?abstractid=6392939&mirid=1&type=2>.
- 27 These are the core targeting obligations under the principles of distinction (API arts 48, 51(2)), precautions (API art. 57) and proportionality under IHL (API arts 51(5)(b), 57(2)(a)(ii)), which are held to the 'reasonable commander standard'.
- 28 Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 12 August 1949, 75 UNTS 31, common art. 1.
- 29 Samantha Besson, *Due Diligence in International Law* (Brill Nijhoff, 2023), p. 23; Bérénice Boutin, 'Between Negligence and Malfunction: How to Address Responsibility for AI Failures', in Bérénice Boutin, Taylor Kate Woodcock and S Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (The Hague: T.M.C. Asser Press, 2026), p. 319.
- 30 Marco Longobardo, 'The Relevance of the Concept of Due Diligence for International Humanitarian Law', *Wisconsin International Law Journal* 37, no. 1 (2020): 44–87, 59–60.
- 31 Bérénice Boutin, 'State Responsibility in Relation to Military Applications of Artificial Intelligence', *Leiden Journal of International Law* 36, no. 1 (2023): 133–150, 143–147; Bérénice Boutin and Taylor Kate Woodcock, 'Aspects of Realizing (Meaningful) Human Control', in Robert Geiß and Henning Lahmann (eds), *Research Handbook on Warfare and Artificial Intelligence* (Cheltenham: Edward Elgar Publishing, 2024), pp. 179–196; Bérénice Boutin, 'Between Negligence and Malfunction: How to Address Responsibility for AI Failures', in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal,*

Ethical, and Technical Dilemmas in Military Artificial Intelligence (The Hague: T.M.C. Asser Press, 2026), p. 320.

- 32 As Copeland highlights in this collected series, article 36 specifically refers to the ‘study, development, acquisition or adoption of a new weapon’ (emphasis added): [insert cross-reference].
- 33 There are compelling cases for why AI systems used for target production should be subject to article 36 review: Justin McClelland, ‘The Review of Weapons in Accordance with Article 36 of Additional Protocol I’, *International Review of the Red Cross* 85, no. 850 (2003): 397–415; Klaudia Klonowska, ‘Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare’, in Terry D Gill, Robin Geiß, Heike Krieger and Rebecca Mignot-Mahdavi (eds), *Yearbook of International Humanitarian Law, Volume 23 (2020)* (Asser Press, 2022) pp. 123–153; Russell Buach, ‘Artificial Intelligence and the Article 36 Legal Review’, *Saint Louis University Law Journal* 70, no. 2 (2026): 285–308, 297–298.
- 34 Grounded for instance in the International Covenant on Civil and Political Rights, December 16, 1966, 999 UNTS 171 (entered into force March 23, 1976).
- 35 Therefore, violations of the core IHL principles will also violate the right to life. See generally *Legality of the Threat or Use of Nuclear Weapons* (Advisory Opinion) [1996] ICJ Rep 226, para. 25; Human Rights Committee (HRC), ‘General Comment No. 36, Article 6: Right to Life’, 30 October 2018, UN Doc. CCPR/C/GC/36, paras 2, 64; Orna Ben-Naftali, ‘Introduction: International Humanitarian Law and International Human Rights Law—Pas de Deux’, in Orna Ben-Naftali (ed.), *International Humanitarian Law and International Human Rights Law* (Oxford: Oxford University Press, 2011), pp. 3–10, 5; Cordula Droegge, ‘The Interplay Between International Humanitarian Law and International Human Rights Law in Situations of Armed Conflict’, *Israel Law Review* 40, no. 2 (2007): 310–355, 310. See also Dieter Fleck, ‘Human Rights in Armed Conflict’, in Dieter Fleck (ed.) *The Handbook of International Humanitarian Law*, 4th edition (OUP, 2021), pp. 449–457, 450.
- 36 HRC, ‘General Comment No 36’ (30 October 2018) UN Doc CCPR/C/GC/36 para. 65.
- 37 Daragh Murray, ‘Adapting a Human Rights-Based Framework to Inform Militaries’ Artificial Intelligence Decision-Making Processes’, *St Louis University Law Journal* 68, no. 2 (2024): 293–326, 304.
- 38 Rachel Davis, ‘The UN Guiding Principles on Business and Human Rights and Conflict-Affected Areas: State Obligations and Business Responsibilities’, *International Review of the Red Cross* 94, no. 887 (2012): 961–979.
- 39 Besson, *Due Diligence*, pp. 101–102.
- 40 Besson highlights that this regulation may even be an obligation in order for states to fulfil their own human rights due diligence obligations: *Due Diligence*, p. 201. Regardless of the potential jurisdictional limitations of IHRL, when read together with IHL it arguably applies to corporate activities that are central to developing a state’s military capabilities in line with the state’s duty to ensure respect for IHL, fulfil core IHL conduct of hostilities duties, and conduct legal reviews.
- 41 ILC Draft Articles on the Responsibility of States for Internationally Wrongful Acts, 2001 YILC, Vol. II (Part Two), 26–30.
- 42 Bérénice Boutin, ‘Between Negligence and Malfunction’, p. 329; Bérénice Boutin, ‘State Responsibility’, pp. 133–150.
- 43 Marta Bo, ‘Are Programmers In or “Out of” Control? The Individual Criminal Responsibility of Programmers of Autonomous Weapons and Self-Driving Cars’, in Sabine Gless and Helena Whalen-Bridge (eds), *Human-Robot Interaction in Law and Its Narratives* (Cambridge University Press, 2024).
- 44 United Nations Human Rights Office of the High Commissioner, *Guiding Principles on Business and Human Rights: Implementing the United Nations ‘Protect, Respect and Remedy’ Framework* (New York and Geneva: United Nations, 2011), Principles 11–24, at: www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinessshr_en.pdf.
- 45 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), *Official Journal of the European Union* L 1689, 12 July 2024, pp. 1–144, at: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.

- 46 Rupert Barrett-Taylor and Matthew Ford, 'Eroding Sovereignty in the Age of "War as a Service"', *Opinio Juris*, 3 November 2025, at: <https://opiniojuris.org/2025/11/03/eroding-sovereignty-in-the-age-of-war-as-a-service/>; 'The Rise of the Forward-Deployed Engineer: Bridging the High-Stakes Chasm Between AI Theory and Execution', *AI Critique*, 20 May 2026, at: www.aicritique.org/us/2026/05/20/the-rise-of-the-forward-deployed-engineer-bridging-the-high-stakes-chasm-between-ai-theory-and-execution/; Alvarez & Marsal, *The Rise and Role of the Forward Deployed Engineer* (Alvarez & Marsal Holdings, 2026), at: www.alvarezandmarsal.com/sites/default/files/2026-04/The%20Rise%20and%20Role%20of%20the%20Forward%20Deployed%20Engineer.pdf; Muzaffer Ünsaldi, 'Defence-as-a-Service: Europe's New Military Procurement Model', *Defence Agenda*, 20 January 2026, at: <https://defenceagenda.com/defence-as-a-service-europe-military-procurement>.
- 47 Katja Bego, *How a Surge in Defence and Dual-Use Technology Investment Could Reconfigure the Global AI Race* (Chatham House, April 2026), at: www.chathamhouse.org/sites/default/files/2026-05/2026-04-28-how-surge-defence-dual-use-investment-could-reconfigure-global-ai-race-bego.pdf.
- 48 Rupert Barrett-Taylor and Natasha Karner, *AI Could Support the Military but Won't Replace the General: Algorithms, Decision-Making and Battlefield Command* (Alan Turing Institute, 2025), at: www.turing.ac.uk/news/ai-could-support-military-wont-replace-general, p. 7; Emily Bienvenue et al., 'Private Tech Companies, the State, and the New Character of War', *Carnegie Endowment for International Peace* (website), 1 December 2025, at: <https://carnegieendowment.org/research/2025/12/ukraine-war-tech-companies>.
- 49 Institute of Electrical and Electronics Engineers (IEEE), *A Framework for Human Decision-Making through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications* (New York NY: IEEE, 2024), at: <https://ieeexplore.ieee.org/stampPDF/getPDF.jsp?arnumber=10707139>.
- 50 On the dual technical and socio-technical lifecycles for military AI capabilities, see Jessica Dorsey et al., 'Bridging Understandings of the Military AI Lifecycle: A Transdisciplinary Socio-Technical Approach to Governance', *Opinio Juris*, 12 June 2026, at: <https://opiniojuris.org/2026/06/12/bridging-understandings-of-the-lifecycle-interdisciplinary-approaches-to-military-ai-governance>.
- 51 See e.g. David N Gibbs, 'The Military-Industrial Complex in a Globalized Context', in Ronald W Cox (ed.), *Corporate Power and Globalization in US Foreign Policy* (New York: Taylor & Francis, 2012), pp. 95–113; Frank Slijper, *The Emerging EU Military-Industrial Complex*, TNI Briefing Series, No. 2005/1 (Amsterdam: Transnational Institute, 2005).
- 52 Barrett-Taylor and Karner, *AI Could Support the Military but Won't Replace the General*, p. 37.
- 53 United Nations, UN Doc. N25/107/66, at: <https://documents.un.org/doc/undoc/gen/n25/107/66/pdf/n2510766.pdf>.
- 54 Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM), *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (The Hague Centre for Strategic Studies, September 2025), at: <https://hcsc.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf>.
- 55 'Regulating Military AI, a Challenge Debated in Geneva Alongside the G7', *Le Monde*, 15 June 2026, at: www.lemonde.fr/en/international/article/2026/06/15/regulating-military-ai-a-challenge-debated-in-geneva-alongside-the-g7_6754479_4.html; 'Tech Executives Attend G7 Summit as Leaders Address AI, Online Safety', *Reuters*, 12 June 2026, at: www.reuters.com/world/tech-executives-attend-g7-summit-leaders-address-ai-online-safety-2026-06-12.
- 56 At the latter, Microsoft was notably the only company present and participating in the discussions: 'Informal Exchanges on Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security—Day 1', *UK Stop Killer Robots*, 16 June 2026, at: <https://ukstopkillerrobots.org.uk/2026/06/16/informal-exchanges-on-artificial-intelligence-in-the-military-domain-and-its-implications-for-international-peace-and-security-day-1>.
- 57 Tshilidzi Marwala, *The Balancing Problem in the Governance of Artificial Intelligence* (Springer, 2024), p. 208.
- 58 Blanchard, Boulanin and Brunn, 'Mapping the Military AI Industry'.
- 59 Vanessa Vohs and Marcel Schliebs, "'This Is My Last Resort": Approaches to Overcome the Stalemate in Autonomous Weapons Regulation Through National Legislation and Industry Self-Regulation', in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (The Hague: T.M.C. Asser Press, 2026).

- 60 Alexander Blanchard, Vincent Boulanin, Laura Brunn and Netta Goussac, 'Dilemmas in the Policy Debate on Autonomous Weapon Systems', *SIPRI* (website), 6 February 2025, at: www.sipri.org/commentary/topical-backgrounders/2025/dilemmas-policy-debate-autonomous-weapon-systems.
- 61 Vohs and Schliebs, "'This Is My Last Resort'"; Marwala, *The Balancing Problem in the Governance of Artificial Intelligence*, p. 208.
- 62 In The Hague in 2023, Palantir's CEO gave an interview on 'his vision on AI and how to use this in the military domain', and the CEO of Avalor AI spoke about 'the role of unmanned vehicles in the military domain, and how Avalor AI is enhancing the tactical value of military unmanned systems' while displaying a demo of such a system: REAIMSummit, 'Alex Karp's (CEO Palantir Technologies) Vision on AI in the Military Domain' (video), *YouTube*, 15 February 2023, at: www.youtube.com/watch?v=L1PnaB15XBI; REAIMSummit, 'Responsible Innovation Talk: Avalor AI' (video), *YouTube*, 16 February 2023, at: <https://www.youtube.com/watch?v=ttXkFEQwnOA>.
- 63 REAIMSummit, '[REAIM Summit 2024] Day 1 _ REAIM Talks' (video), *YouTube*, 9 September 2024, at: www.youtube.com/watch?v=zUY_Wh9yZsk. Also in Seoul, the CEO of Two Platforms, a startup focused on 'artificial reality' and generative AI, sat on a panel discussion alongside policymakers, academics and members of other international institutions for a plenary session on 'Envisioning the Future Governance of AI in the Military Domain': REAIMSummit, '[REAIM Summit 2024] Day 2 _ Plenary Session 3' (video), *YouTube*, 10 September 2024, at: www.youtube.com/watch?v=3_QCOrZrRbc.
- 64 Many industry booths were present with representatives from companies such as Microsoft, Capgemini, Oracle and Proportione, to name just a few.
- 65 *REAIM Summit 2026: Program* (Madrid: Ministry of Foreign Affairs, European Union and Cooperation, 2026), at: https://ream2026.es/es/REAM2026/Documents/Program%20REAM%202026_20260129.pdf.
- 66 Tim Sweijs and Sofia Romansky, *International Norms Development and AI in the Military Domain*, CIGI Papers, No. 300 (Waterloo ON: Centre for International Governance Innovation, 2024).
- 67 Schwarz, 'From Blitzkrieg', p. 13.
- 68 Amos Toh and Emile Ayoub, *The Business of Military AI* (Brennan Center for Justice, 2026), at: www.brennancenter.org/media/15340/download/bcj-167_business_of_military_ai_final.pdf.
- 69 This has been done in particular by Palantir, including at the REAIM Summit in Seoul in 2024, as well as when addressing the United Kingdom House of Lords on autonomous weapons: Palantir Technologies UK, Ltd, *Submission to the House of Lords AI in Weapons Systems Committee: Inquiry on AI in Weapons Systems* (Palantir, April 2023), at: www.palantir.com/assets/xrfr7uokpv1b/T6XBvRNbtgOys-f4XuYqpH/1ba005ae5b469eb47ea484ad34dadcea/Palantir_Submission_to_the_HL_AI_in_Weapons_Systems_Committee.pdf.
- 70 Though collaborators on this program, led by UNIDIR with the support of the OHCHR, are not publicly listed, known industry actors involved include Microsoft, Hitachi America Ltd and the Japan Defense Technology Foundation: 'Framework of Responsible Industry Behaviour for AI in the Military Domain', *UNDIR* (website), at: <https://unidir.org/framework-of-responsible-industry-behaviour-for-ai-in-the-military-domain>.
- 71 Ibid.
- 72 Barrett-Taylor and Karner, *AI Could Support the Military but Won't Replace the General*, pp. 7, 9.
- 73 Woodcock, 'Human-Machine (Learning) Interactions'; Dorsey, 'The Erosion of Human(e) Judgement in Targeting?'. Barrett-Taylor situates contemporary developments for AI in warfare in the historical trend of militaries using quantification and data processing to manage complexity since at least the 1950s: Rupert Barrett-Taylor and Gavin Wilde, 'A Digitized, Efficient Model of War', *Carnegie Endowment for International Peace* (website), 3 June 2025, at: <https://carnegieendowment.org/research/2025/06/a-digitized-efficient-model-of-war>.
- 74 Renic and Schwarz 'Crimes of Dispassion', pp. 337–338. Spaulding highlights that 'the displacement of human judgment and the deterioration of conditions for its exercise ... are signal attributes of bureaucratic systems': Norman W Spaulding, 'Is Human Judgment Necessary? Artificial Intelligence, Algorithmic Governance, and the Law', in Markus D Dubber et al. (eds), *The Oxford Handbook of Ethics of AI* (Oxford: Oxford University Press, 2020), p. 392.

- 75 See e.g. Marco Sassòli, 'Targeting: The Scope and Utility of the Concept of "Military Objectives" for the Protection of Civilians in Contemporary Armed Conflicts', in David Wippman and Matthew Evangelista (eds), *New Wars, New Laws? Applying the Laws of War in 21st Century Conflicts* (Transnational Publishers, 2005), pp. 181–210, 204; Amanda Alexander, 'A Short History of International Humanitarian Law', *European Journal of International Law* 26, no. 1 (2015): 109–138.
- 76 Klonowska and Woodcock, 'Rhetoric and Regulation'; Rehak and Woodcock, 'Automating Civilian Harm'; Lahmann, 'Bad Algorithms'.
- 77 'How Tech Giants Turned Ukraine Into an AI War Lab', *TIME*, at: <https://time.com/6691662/ai-ukraine-war-palantir>.
- 78 Taylor Kate Woodcock, 'Human/Machine(-Learning) Interactions, Human Agency and the International Humanitarian Law Proportionality Standard', *Global Society*, 38, no 1 (2024): 100–121; Woodcock, 'Human-Machine (Learning) Interactions'.
- 79 Yuval Abraham, "'Lavender": The AI Machine Directing Israel's Bombing Spree in Gaza', *+972 Magazine*, 3 April 2024, at: www.972mag.com/lavender-ai-israeli-army-gaza.
- 80 Rehak and Woodcock, 'Automating Civilian Harm'.
- 81 Woodcock, 'Human-Machine (Learning) Interactions'.
- 82 Zeerak Waseem et al., 'Disembodied Machine Learning: On the Illusion of objectivity in NLP', *arXiv*, 2101.11974 (2021), at: <https://arxiv.org/abs/2101.11974v1>.
- 83 Cummings and Li, 'Subjectivity in the Creation of Machine Learning', pp. 2–3. These include: 'Picking the modeling approach to be used, Picking which features should be included out of large datasets, Determining whether to drop cases with missing data or to generate missing data estimates, Picking a p value for statistical significance, Deciding numbers of neurons for hidden NN layers, Picking the maximal training iteration for the NN training process, Picking stopping rules for training performance factors (software-dependent), Selecting data training and testing ratios, Picking threshold values between binary unbalanced outcomes, Choosing thresholds between important/unimportant features, Determining whether model accuracy is good enough, Deciding what the actual important features for a model' (ibid., p. 14).
- 84 Laurence E Diver, *Digisprudence: Code as Law Rebooted* (Edinburgh University Press, 2022), 5
- 85 Antoine Smallegange, Jurriaan van Diggelen and Jonathan Kwik, 'Characteristics of the Military Domain and Their Impact on Military AI Applications', in *2025 International Conference on Military Communication and Information Systems (ICMCIS 2025): Oeiras, Portugal, 13–14 May 2025* (Piscataway NJ: IEEE, 2025), pp. 232–241, at: <https://doi.org/10.1109/ICMCIS64378.2025.11047950>.
- 86 Zachary C Lipton, 'The Mythos of Model Interpretability', *arXiv*, 1606.03490 (2016): 5, at: <https://arxiv.org/abs/1606.03490>, Riccardo Guidotti et al., 'A Survey of Methods for Explaining Black Box Models', *arXiv*, 1802.01933 (2018): 1, 2, 5, 11, at: <https://arxiv.org/abs/1802.01933>; Brent Mittelstadt et al., 'Explaining Explanations in AI', *arXiv*, 1811.01439 (2019): 2, at: <https://arxiv.org/abs/1811.01439>.
- 87 Woodcock, 'Machine Learning: AI, Algorithmic Arbitrariness and Legal Decisions'.
- 88 Matteo Pasquinelli and Vladan Joler, 'The Noosphere Manifested: AI as Instrument of Knowledge Extractivism', *AI & Society* 36 (2021): 1263–1280, 1278; Waseem et al, 'Disembodied Machine Learning', p. 1.
- 89 Meltem Bostanci and Özgür Yilmaz, 'Labour and Exploitation Processes in Artificial Intelligence: Example of Digital Taylorism in Data Labelling', *Journal of Social & Cultural Studies* 15 (2025): 24–48; Julian Posada, 'Embedded Reproduction in Platform Data Work', *Information, Communication & Society* 25, no. 6 (2022): 816–834.
- 90 Pasquinelli and Joler, 'The Noosphere Manifested'; Milagros Miceli, Martin Schuessler and Tianling Yang, 'Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision', *arXiv*, 2007.14886 (2020): 2, at: <https://arxiv.org/abs/2007.14886>.
- 91 Bhavya Ghai, Q Vera Liao, Yunfeng Zhang and Klaus Mueller, 'Measuring Social Biases of Crowd Workers Using Counterfactual Queries', *arXiv*, 2004.02028 (2020), at: <https://arxiv.org/abs/2004.02028>; Christoph Hube, Besnik Fetahu and Ujwal Gadiraj, 'Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2019), 407: 1–12, at: <https://doi.org/10.1145/3290605.3300637>; Fabian L Wauthier and Michael I Jordan, 'Bayesian Bias Mitigation for

- Crowdsourcing', in *Advances in Neural Information Processing Systems 24* (NIPS, 2011), at: <https://people.eecs.berkeley.edu/~jordan/papers/wauthier-jordan-nips12.pdf>.
- 92 danah boyd and Kate Crawford, 'Six Provocations for Big Data', in *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011* (SSRN, 2011), p. 5, at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.
- 93 Cummings and Li, 'Subjectivity in the Creation of Machine Learning', 3.
- 94 Gary Marcus, 'The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence', *arXiv*, 2002.06177 (2020): 5–6, 16.
- 95 Neil Fraser, 'Neural Network Follies', *Neil Fraser* (website), September 1998, at: <https://neil.fraser.name/writing/tank> (accessed 5 July 2026).
- 96 Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016); Alex A Freitas, 'Comprehensible Classification Models—A Position Paper', *ACM SIGKDD Explorations Newsletter* 15, no. 1 (2014): 2.
- 97 Woodcock, 'Human-Machine (Learning) Interactions'.
- 98 Ibid.
- 99 Mireille Hildebrandt, 'Data-Driven Prediction of Judgment: Law's New Mode of Existence?', in *Collected Courses of the Academy of European Law* (Oxford: Oxford University Press, 2019), EUI Summer School, p. 6, at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3548504.
- 100 Concerns around bias in targeting, such as the targeting of military-aged men from various regions as militia members, pre-dates the use of AI systems: see Kristina Benson, 'Kill 'em and Sort It Out Later: Signature Drone Strikes and International Humanitarian Law', *Pacific McGeorge Global Business & Development Law Journal* 27, no 1 (2014): 17–51, 34–36.
- 101 Woodcock, 'Human-Machine (Learning) Interactions'.
- 102 Netta Goussac and Vincent Boulanin, *Responsible Procurement of Military Artificial Intelligence* (SIPRI, 2026), pp. 6–7, at: www.sipri.org/sites/default/files/2026-02/0226_milai_procurement_260216.pdf.
- 103 An example is in the United States, which implemented the Better Buying Power Directive and Tradewinds initiative to foster more efficient military acquisition: Amaani Lyle, 'Better Buying Power 3.0 Stresses Innovation, Affordability', *Defense Logistics Agency* (website), 15 April 2015, at: www.dla.mil/About-DLA/News/News-Article-View/Article/633556/better-buying-power-30-stresses-innovation-affordability; *Tradewinds* (website), at: www.tradewindai.com.
- 104 [Insert cross-reference to Manoj Harjani's special issue paper on TEVV].
- 105 Alan Backstrom and Ian Henderson, 'New Capabilities in Warfare: An Overview of Contemporary Technological Developments and the Associated Legal and Engineering Issues in Article 36 Weapons Reviews', *International Review of the Red Cross* 94, no 886 (2012): 483–514, 508.
- 106 Ibid.
- 107 Ibid., p. 509, suggesting that the transposition of legal requirements for the purposes of review should be 'testable, quantifiable, measurable, and reasonable'.
- 108 Marcus, 'The Next Decade in AI', p. 39; François Chollet, 'On the Measure of Intelligence', *arXiv*, 1911.01547 (2019): 3.
- 109 Missy L Cummings, 'Artificial Intelligence and the Future of War', in ML Cummings et al. (eds) *Artificial Intelligence and International Affairs: Disruption Anticipated* (London: Chatham House, 2018), at: <http://chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>, p. 12.
- 110 Peter Layton, *Fighting Artificial Intelligence Battles: Operational Concepts for Future AI-Enabled War*, Joint Studies Paper Series No. 4 (Canberra: Department of Defence, 2021), doi:10.51174/JPS.004, p. 14.
- 111 Martin Hagström, 'Military Applications of Machine Learning and Autonomous Systems', in Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Volume I Euro-Atlantic Perspectives* (SIPRI, 2019), pp. 32–38, 37; International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (Geneva: ICRC, 2019), p. 3, at: www.icrc.org/sites/default/files/document/file_list/autonomy_artificial_intelligence_and_robotics.pdf. Boulanin points to the lack of reliable methods for testing and evaluating the performance of

- systems that are 'complex, adaptive and nonlinear': Vincent Boulanin, *Implementing Article 36 Weapon Reviews in the Light of Increasing Autonomy in Weapon Systems*, SIPRI Insights on Peace and Security, No. 2015/1 (SIPRI, 2015), p. 15.
- 112 Hagström, 'Military Applications of Machine Learning', pp. 37–38.
- 113 Brian K Hall, 'Autonomous Weapons Systems Safety', *Joint Force Quarterly* 86 (2017): 86–93, 89.
- 114 Boulanin, *Implementing Article 36*, p. 15.
- 115 This can include what has been referred to as 'user-specific' explainable AI failures: Clara Bove et al., 'Why Do Explanations Fail? A Typology and Discussion on Failures in XAI', *arXiv*, 2405.13474v1 (2024): 7–9, at: <https://arxiv.org/html/2405.13474v1>.
- 116 ICRC, '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', *International Review of the Red Cross* 88, no 864 (2006): 937.
- 117 Yves Sandoz, Christophe Swinarski and Bruno Zimmermann (eds), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (Geneva: ICRC, 1987), para. 1473.
- 118 McClelland, 'The Review of Weapons', p. 413.
- 119 Woodcock, 'Human-Machine (Learning) Interactions'.
- 120 Klonowska and Woodcock 'Rhetoric and Regulation'.
- 121 Woodcock, 'Human-Machine (Learning) Interactions'.
- 122 Woodcock, 'Human/Machine(-Learning) Interactions, Human Agency and the International Humanitarian Law Proportionality Standard'.
- 123 Boutin, 'State Responsibility', p. 146.
- 124 This is acknowledged by the range of 'lifecycle' approaches to military AI and the important work being done on compliance-, lawfulness- and legal protection-by-design for AI systems. See e.g. GC REAIM, *Responsible by Design*; Boutin and Woodcock, 'Legal Aspects'; 'Lawful by Design Initiative', *Article 36 Legal*, at: www.article36legal.com/lawful-by-design; Mireille Hildebrandt, 'Saved by Design? The Case of Legal Protection by Design', *Nanoethics* 11 (2017): 307–311, at: <https://doi.org/10.1007/s11569-017-0299-0>.
- 125 Goussac and Boulanin, *Responsible Procurement of Military Artificial Intelligence*, p. 19.
- 126 Ibid, p. 18.
- 127 Backstrom and Henderson, 'New Capabilities in Warfare', p. 513, suggesting that each discipline should 'have enough understanding of the other fields to appreciate potential interactions, facilitate meaningful discussion, and understand their own decisions in the context of impacts on other areas of development'.
- 128 Lena Trabucco, *International Humanitarian Law and Lethal Autonomous Weapon Systems* (Centre for Military Studies, University of Copenhagen, 2024), at: https://cms.polsci.ku.dk/publikationer/folkeretten-og-doeedbringende-autonome-vaabensystemer/International_Humanitarian_Law_and_Lethal_Autonomous_Weapon_Systems.pdf.
- 129 Trevor Taylor, 'Artificial Intelligence in Defence: When AI Meets Defence Acquisition Processes and Behaviours', *RUSI Journal* 164, nos 5–6 (2019): 72–81, 76.
- 130 Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Pearson, 2022), pp. 669–670.
- 131 Ibid.
- 132 See e.g. Barrett-Taylor and Ford, 'Eroding Sovereignty'.
- 133 Klaudia Klonowska, *War, AI Decision-Support Systems, and International Humanitarian Law* (Routledge, 2027). In another publication, Klonowska highlights some of the subjectivities that become embedded into AI systems through 'tinkering' processes resulting in feedback loops, including performance monitoring and model retraining that use methods involving user feedback for offline learning systems. She highlights that maintaining the performance and reliability of such systems involves consistent redevelopment, creating challenges not only for ensuring the lawfulness of military operations based on the reliability of these systems but also for conducting meaningful review of the lawfulness of these systems before they result in harm: Klaudia Klonowska, 'Elusive Reliability: Techno-Legal Considerations on the Challenges of Maintaining and Reviewing Performance of AI in Military Applications', *Computer Law & Security Review* 6, no. 1063181 (2026): 1–22, 6.

- 134 Jun Wu, 'Distribution Shifts in Trustworthy Machine Learning', *AI Magazine* 47 (2026): 1–15, at: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.70057>, p. 1.
- 135 See e.g. Sophia Sheera, 'US Military Used Musk's Grok AI to Fire 2,000 Missiles at Iran: Court Documents Reveal the Trillionaire's Role in Operation Epic Fury', *Novara Media*, 18 June 2026, at: <https://novaramedia.com/2026/06/18/us-military-used-musks-grok-ai-to-fire-2000-missiles-at-iran>; Josh Taylor, 'Pokémon Go Data Trained AI That Could Assist Military Drones in War Zones', *The Guardian*, 12 June 2026, at: www.theguardian.com/technology/2026/jun/12/pokemon-go-data-trained-ai-that-could-assist-military-drones-in-war-zones.
- 136 Myra Cheng et al., 'Social Sycophancy: A Broader Understanding of LLM Sycophancy', *arXiv*, 2505.13995 (2025); Jonathan Kwik, 'Digital Yes-Men: How to Deal with Sycophantic Military AI?', *Global Policy* 16, no. 3 (2025): 467–473.
- 137 Wu, 'Distribution Shifts', p. 1.
- 138 Keeping in mind that violations for unintended engagements will turn on the reasonableness of commanders' decisions, this raises the question of whether reasonable system engineering choices are also part of this calculation, or the 'reasonable military software engineer'.
- 139 See e.g. Damian Copeland and Luke Reynoldson, 'How to Avoid "Summoning the Demon": The Weapons Review of Weapons with Artificial Intelligence', *Pandora's Box* 97 (2017): 106–108; Damian Copeland, *A Functional Approach to the Legal Review of Autonomous Weapon Systems*, International Humanitarian Law Series, Vol. 72 (Brill, 2025), p. 182; Netta Goussac et al., *Enhancing the Legal Review of Autonomous Weapon Systems: Report of an Expert Meeting* (Brisbane: TC Beirne School of Law, The University of Queensland, 2023), doi:10.14264/2bbfd31; Jonathan Kwik, 'Iterative Assessment for Military Artificial Intelligence (AI) Systems', in Bérénice Boutin, TK Woodcock and S Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence* (The Hague: T.M.C. Asser Press, 2026); Klaudia Klonowska and Jonathan Kwik, 'Artificial Intelligence in Contemporary Conflicts and the Future of Military Law', in Paul AL Ducheine, Terry D Gill, Peter BMJ Pijpers and Marten C Zwanenburg (eds), *A Research Agenda for Military Law* (Edward Elgar, 2026), pp. 107–108.
- 140 Alex Krasodonski, 'Anthropic's Feud with the Pentagon Reveals the Limits of AI Governance', *Chatham House*, 4 March 2026, at: www.chathamhouse.org/2026/03/anthropics-feud-pentagon-reveals-limits-ai-governance.
- 141 Erin Griffith, 'Google Won't Renew Controversial Pentagon AI Project', *Wired*, 1 June 2018, at: www.wired.com/story/google-wont-renew-controversial-pentagon-ai-project.
- 142 'Palantir Expands Maven Smart System AI/ML Capabilities to Military Services', *Palantir* (website), 20 September 2024, at: <https://investors.palantir.com/news-details/2024/Palantir-Expands-Maven-Smart-System-AI/ML-Capabilities-to-Military-Services>.



researchcentre.army.gov.au