Asia Pacific Journal of International Humanitarian Law

Artificial Intelligence and Warfare

Special Issue
2026 Edition

ASIA-PACIFIC JOURNAL OF INTERNATIONAL HUMANITARIAN LAW

ARTIFICIAL INTELLIGENCE AND WARFARE

SPECIAL EDITION | 2026

Philippine Copyright © 2026

by University of the Philippines Law Complex and
International Committee of the Red Cross

Published by U.P. LAW COMPLEX

ISSN: 2719-1141 (online)
ISSN: 2799-1725 (print)

No part of this book may be reproduced in any form, or by any electronic or mechanical means, including information storage and retrieval systems, without permission in writing from the author and publisher, except by a reviewer who may quote brief passages in a review.

ALL RIGHTS RESERVED BY THE AUTHORS

CURRENT BOARD OF EXPERTS

Prof. Alberto Costi
University of Wellington

Dr. Borhan Uddin Khan
University of Dhaka

Dr. Jonathan Kwik
University of Amsterdam

Dr. Marnie Lloydd
Victoria University of Wellington

Dr. Jan Römer
International Committee of the Red Cross

Prof. Joon-Koo Yoo
IFANS (Institute of Foreign Affairs and Security)

Dr. Papawadee Tanodomdej
Chulalongkorn University

Dr. Ai Kihara-Hunt
University of Tokyo

Dr. He Tiantian
Chinese Academy of Social Sciences

Judge Raul C. Pangalangan
University of the Philippines

Prof. Deepika Udagama
University of Peradeniya

Mx. Ioanna Voudori
International Committee of the Red Cross

Mx. Tamalin Bolus
International Committee of the Red Cross

FORMER BOARD OF EXPERTS

Dr. Monique Cormier
Monash University

Mr. Milan Jovančević
United Nations Assistance to the Khmer Rouge Trials (UNAKRT)

Prof. Junaenah Sulehan
University College Sabah Foundation

Prof. Kyo Arai
Doshisha University (Kyoto)

Dr. Matthias Vanhullebusch
Shanghai Jiaotong University

Ms. Kelisiana Thynne
International Committee of the Red Cross

Dr. Mohd Hisham Mohd Kamal
International Islamic University Malaysia

Prof. Nohyoung Park
Korea University

Dr. Liu Guofu
Beijing Institute of Technology

Ms. Georgia Hinds
International Committee of the Red Cross

Sahar Haroon
International Committee of the Red Cross

EDITORIAL TEAM

Prof. Rommel J. Casis
Managing Editor
University of the Philippines

Atty. Joan Paula Deveraturda
Associate Editor

Ms. Aira Lynn Cunanan
Senior Editorial Assistant

Mr. Mackie Valenzuela
Junior Editorial Assistant

Mr. John Paolo V. Cabrera
Junior Editorial Assistant

Ms. Patricia Isabel G. Arago
Junior Editorial Assistant

Ms. Ma. Flordeliza Villar
Layout Artist

Mr. Raphael Teodoro Regala
Copy Editor

INTRODUCTION: BACKGROUND AND OVERARCHING THEMES

Samuel White*
Jonathan Kwik**
Danielle Yeow***
Jon Truby****

Background to the Special Issue

For much of modern military history, advances in technology have altered how wars are fought without fundamentally displacing the human decision-maker from the centre of military judgment. The machine gun increased lethality, the aircraft expanded reach, whilst the computer accelerated the processing of information; yet the assessment of military necessity, the weighing of civilian harm and the exercise of legal and moral responsibility remained recognisably human functions. Artificial intelligence (AI) presents a different challenge. Rather than merely extending human capabilities, AI systems increasingly participate in the production of military knowledge itself. They identify patterns, generate recommendations, prioritise targets, allocate resources, and shape the information environment through processes that are often opaque, adaptive, and difficult even for their designers to fully explain. As these systems become integrated into military planning and operations, longstanding assumptions about control, accountability, and legal responsibility are placed under increasing pressure.

This transformation is occurring at a moment when the character of conflict is itself changing (although historians would say, correctly, that the character of conflict is always changing). Contemporary warfare is increasingly data-driven, digitally enabled and fought across interconnected physical, cyber, cognitive, and informational domains. The strategic value of data, algorithms, cloud infrastructure, and computational power now rivals that of many traditional military assets. Military advantage is no longer derived solely from the destruction of an adversary's forces, but also from the ability to collect, process, manipulate, and act upon

* Scientia Senior Lecturer in Military Law and War Studies, UNSW Canberra; Global Fellow at the Centre for International Law, National University of Singapore.
** Researcher in International Law, TMC Asserinstituut (The Hague); Drafting Team, AI Manual Project; Board of Experts member, Asia-Pacific Journal of International Humanitarian Law.
*** Adjunct Senior Research Fellow and Lead, Cyber Law and Governance at the Centre for International Law, National University of Singapore.
**** UNESCO Chair on Artificial Intelligence Law & Sustainability at the Centre for International Law, National University of Singapore.

information at speed and scale. Artificial intelligence sits at the centre of this shift. It promises unprecedented operational efficiencies and decision superiority, while simultaneously generating novel risks that challenge existing legal frameworks and ethical understandings of warfare. Questions of accountability, reliability, transparency, autonomy, and human control are no longer speculative concerns. They have become practical issues confronting military planners, policymakers, and legal advisers in real-world operations.

Recent conflicts have brought these questions into sharper focus, in particular the recent conflict between the United States of America (and its allies) against Iran. The integration of AI-enabled decision-support systems into targeting processes, intelligence analysis, and military planning has demonstrated both the operational potential of these technologies and the difficulties they pose for existing legal regimes. At the same time, the growing strategic importance of data infrastructure, cloud computing services, and commercial technology providers has further blurred traditional distinctions between civilian and military spheres. What was once viewed as a future challenge is now an immediate reality. The law is increasingly required to address not merely autonomous weapons, but an entire ecosystem of AI-enabled military capabilities whose effects extend across the full spectrum of conflict.

AI systems aggregated and processed drone feeds, satellite imagery, signals intelligence, and intercepted communications at speeds beyond human capacity, producing strike recommendations that effectively collapsed what had once been layered, time-consuming approval chains into near-real-time outputs.[1]

Iran's response has, in turn, focused on disrupting the data architecture that enables this form of warfare. In March and April 2026, Iranian forces targeted Amazon Web Services (AWS) data centres in the UAE and Bahrain using drones and missiles, marking the first known instance of a state deliberately striking commercial data infrastructure during an armed conflict.[2] The resulting outages,

[1] Hope Hodge Seck, "AI Tool Has 'Saved a Lot of Aircraft' in Epic Fury, AFSOC Chief Says", *Military Times*, 13 May 2026, available at: https://www.militarytimes.com/news/ your-military/2026/05/13/ai-tool-has-saved-a-lot-of-aircraft-in-epic-fury-afsoc-chief-says/ accessed 14 July 2026

[2] Shubham Kaila, Aditya Soni, and Mrinmay Dey, "Iranian Strikes Target AWS Facilities in Bahrain and UAE", *Reuters*, 3 March 2026, available at: https://www.reuters.com/ world/middle-east/amazon-cloud-unit-flags-issues-bahrain-uae-data-centers-amid-iran-strikes-2026-03-02/. accessed 14 July 2026

connectivity failures, and fire damage illustrate how civilian digital systems have become integral to military operations.

It is against this backdrop that this Special Issue is situated. This conflict did not emerge as an isolated technological departure. Rather, it builds on a trajectory visible in earlier uses of AI in the Gaza Strip, where Israeli systems such as "Lavender" and "Gospel" reportedly generated tens of thousands of targeting recommendations from surveillance data which has generated in its wake, concerns about the risks posed by the deployment of such technology[3] and the erosion of civilian protection.[4] It also extends developments associated with Project Maven, the United States military's primary machine learning initiative, which has supported targeting operations across Iraq, Syria, Yemen, and the Red Sea since 2024. At the same time, these developments have sharpened tensions between the ethical positions of commercial AI developers and the operational requirements of military institutions. On the day Operation Epic Fury commenced, Anthropic was reportedly placed on a Pentagon security blacklist after declining to provide unrestricted access to its Claude system, despite its prior integration into strike planning through Project Maven and classified AWS environments.[5]

Taken together, these developments mark a decisive moment. Debates surrounding the ethics, legality, and governance of military AI have moved beyond abstraction and into direct operational practice. What had previously been framed as a prospective challenge is now an immediate and tangible reality, unfolding in real time across interconnected domains of warfare, infrastructure, and commerce.

This Special Issue proceeds from that premise. It brings together a set of contributions to a regional roundtable Workshop in Singapore on 16-17 April 2026, hosted by the National University of Singapore's Centre for International Law and in support of the wider initiative of the *West Point Manual on International Law*

---

[3] Human Rights Watch, Questions and Answers: Israeli Military's Use of Digital Tools in Gaza, 10 September 2024, available at https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-of-digital-tools-in-gaza#_On_which_grounds. accessed 14 July 2026

[4] Emelie Andersin, "Military Use of Biometrics series – Isreal's Use of AI-DSS and Facial Recognition Technology: The Erosion of Civil Protection in Gaza", *Articles of War – Blog of the Lieber Institute* , 24 October 2025, available at: https://apjihl.org/wp-content/uploads/2026/01/APJIHL-Guidelines-for-Authors.pdf. accessed 14 July 2026

[5] Hayden Field, "Anthropic and U.S. Government to Face Off in D.C. Court Over Blacklisting of AI Company", *CNBC*, 19 May 2026, available at: https://www.cnbc.com/2026/05/19/anthropic-dod-blacklist-court-opening-arguments.htm. accessed 14 July 2026

Applicable to Artificial Intelligence in Warfare ('AI Manual').[6] The papers herein were selected for their novelty and deep understanding of how to engage with the *sui generis* tension that AI brings. To that end, we seek to address issues that find no precedent, such as data poisoning, AI self-discovering perfidious tactics, agency spillovers, and regional data inadequacies. The articles in this issue examine, from a range of disciplinary perspectives, how established frameworks of international humanitarian law (IHL), state responsibility, and military practice are being strained, adapted, or reconstituted in response to these developments. In doing so, the issue seeks not only to diagnose the present moment, but to clarify the contours of the challenges that lie ahead.

Prominent Themes in the Contributions

In discussing the application of the law to any novel technology in warfare, a crucial question invariably arises: What is truly new, and what is merely an amplification or evolution of an existing problem? It is a core question regularly posed by members of the AI Manual's drafting team, and often not one easily answered. Autonomy, for example, is frequently associated with being a unique characteristic of AI systems, but has been a feature of cyber-capabilities for a long time.[7] In contrast, other properties of such systems—such as their stochasticity,[8] opacity,[9] unintuitive failure modes[10] or the ability to creatively find solutions[11]—are unique to AI, and accordingly give rise to novel legal problems and potential legal gaps. Identifying these risks in the first place may be as difficult as mitigating them, often requiring in-depth knowledge of the technology and the setting in which they are to be deployed (military operations).

In this Special Issue, several contributors have identified, analysed and attempted to address *AI-specific challenges and how they impact the application of international law*. Dora Velenczei theorises how goal-based optimisers in AI-enabled

[6] Centre for International Law, National University of Singapore, *West Point AI Manual Workshop Event Report*, Singapore, 16–17 April 2026, available at: https://cil.nus.edu.sg/wp-content/uploads/2026/05/AI-Manual-Workshop-Event-Report.pdf.
[7] Laurent Gisel and Lukasz Olejnik, *The Potential Human Cost of Cyber Operations*, ICRC Expert Meeting 14–16 November 2018, ICRC, 2018, p. 38.
[8] Nathan G. Wood *et al.*, "Stop Saying 'AI'", arXiv, 25 February 2026, p. 23.
[9] Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI*, UNIDIR, Geneva, Switzerland, 22 September 2020.
[10] See Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies*, TMC Asser Press, The Hague, 2024, pp. 114–120.
[11] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Pearson, London, 2021, pp. 53–54.

weapon systems may discover methods to fool enemies by (unlawfully) misrepresenting themselves as protected entities—potentially in violation of the prohibition on perfidy.[12] Samuel White and Daniel Harker further raise the dangers of data poisoning (a tactic where adversaries exploit a vulnerability unique to AI, by compromising the AI system as it is being trained)[13] and how these may cause erratic or dangerous behaviours in the affected AI systems, possibly in violation of international law. Anviksha Pachori raises in her contribution the problem of AI's continued evolution through updates, modifications and patches,[14] which generally conflicts with obligations under IHL that often assume static, relatively unchanging capabilities, and suggests that a legal principle of 'IHL-by-Design' can potentially address those tensions.

Fears of AI systems overstepping their boundaries, amongst others due to the issues described above, have led to several propositions to *'code in' legal or ethical limitations*.[15] Sze-Hong Lam, however, strongly questions in his contribution whether such proposals are feasible. Through an in-depth analysis of various sources of indeterminacies that lie at the very core of IHL, Lam argues that integrating the laws of war into AI only succeeds in a somewhat utopian vision of IHL that does not comport to realities. Indeed, if coding in legal limitations directly is an unreliable way to ensure legal compliance, it falls to the users and developers of the AI systems to ensure that this is the case. Through his *dignitary design* concept, Sergey Sayapin advances a comprehensive, lifecycle-spanning framework for embedding respect for human dignity throughout the design, governance, policy, and deployment practices of military AI systems.

Multiple contributions demonstrate that obligations that are the most evident to IHL-lawyers as ways to constrain conduct during warfare—such as the rules related to proportionality and precautions in Art. 57(2) of Additional Protocol I—may not be available to address AI-specific problems. Many use-cases of AI such as

---

[12] For a similar argument on how this may be possible with AI-based cyber defence agents, see also Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Lieber Institute West Point*, 5 December 2025.

[13] See Micah Goldblum *et al.*, "Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses", 18 December 2020.

[14] Klaudia Klonowska, "Techno-Legal Tinkering in War: AI Decision-Support Systems and International Humanitarian Law", Doctoral thesis, University of Amsterdam, 18 March 2026, pp. 26–27.

[15] E.g. Ronald C. Arkin, Patrick Ulam, and Brittany Duncan, *An Ethical Governor for Constraining Lethal Action in an Autonomous System*, Georgia Institute of Technology Atlanta Mobile Robot Lab, 2019.

decision-support systems (AI-DSS) fall below the threshold of an attack, and poisoning attacks such as those discussed by White and Harker are likely to be used as a grey zone tactic, and executed before an armed conflict even commences.[16] As a result, many authors also take to analysing *complementary legal rules and instruments* as a means to regulate military AI systems. Pachori's contribution which focuses on Additional Protocol I Article 36 reviews and the Arms Trade Treaty, and the Martens Clause has received particular attention in this Issue, with White and Harker, Pachori, and Sayapin all dedicating time to analyse whether and how the "principles of humanity and the dictates of public conscience" can be operationalised to drive the protection of civilians when AI systems are involved.

In addition to contributions exploring how existing law can be applied to military AI systems, the possibility and feasibility of *new governance and norm or treaty-building* has also been explored extensively in this Issue. Zhixiong Huang and Haokun Yu in this regard advance a dual argument: that existing law is insufficient to properly regulate military AI systems ('de jure necessity'), but that several prevailing factors render the possibility of a comprehensive treaty that also binds the Great Powers very unlikely ('de facto improbability'). They nevertheless end with a constructive note, noting the normative benefit of nevertheless pursuing the treaty-making process in the longer term. Bagus Jatmiko echoes this view, also encouraging international engagement and discourse on military AI as an important initiative that States, and in particular middle-power States such as those in Southeast Asia, should pursue.

Mei-Ching Liu, however, issues a word of warning on how challenging such initiatives may prove in practice. In her contribution, she extensively documents and analyses the discourse at the Group of Governmental Experts (GGE) in Geneva surrounding the singular word 'lethal' in LAWS (Lethal Autonomous Weapon Systems) between 2013 and 2026. From her exploration on the way States have used, developed and occasionally instrumentalised the notion of lethality, she highlights how even small semantic or conceptual disagreements in the diplomatic process can entail long-term legal and practical consequences.

[16] See Michael N. Schmitt, "Big Data: International Law Issues Below the Armed Conflict Threshold", in L.A. Dickinson and E.W. Berg (eds.), *Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold*, Oxford University Press, 14 March 2024.

Challenges in *ensuring legal accountability* for the use of military AI systems has been a dominant concern for more than a decade of international discourse.[17] Fan Yang theorises that the phenomenon of 'agency spillover', where machine agency deviates from human presets, lies at the root of difficulties to assign accountability under multiple legal regimes, including IHL, State responsibility and individual criminal liability. Yang nevertheless adds a constructive leg to his critique by recommending that the law should instead consider human-machine hybrid agency as the paradigm when assessing legal accountability for AI systems, which overcome many obstacles he identified. Drexx Laggui, subsequently, adds a practical perspective by considering how States and victims may practically collect and audit proof of misconduct by military AI. He argues that AI systems are prone to leaving digital evidence trails and records that, given the proper forensic procedures and mechanisms, can be obtained and safeguarded for use in legal proceedings.

Finally, the NUS Workshop was also conceived as an opportunity to discuss problems and concerns arising from the development and use of military AI *from an Asia-Pacific and Southeast Asian lens*—a region that, in a field of discourse mostly examined through Western and Chinese perspectives, has received relatively little scholarly attention.[18] Erick Nielson Javier in his contribution conducted a thorough analysis of the state of military AI in the Southeast Asian region, which features unique strategic and security challenges. Looking at both the national and regional interests of Southeast Asian States in military AI, but also shortcomings in their national strategies or acquisition processes, he argues that these States need to seriously consider and consolidate what ways forward they wish to pursue. Finally, in advancing a similar argument, Bagus Jatmiko juxtaposed several problems unique to AI to the Southeast Asian region. He argues that a lack of regional-representative datasets and other factors (such as opacity and human-machine dynamics like sycophancy),[19] which may exacerbate the lower technical expertise of prospective users, could raise novel risks if AI are integrated carelessly into the region, and recommends active engagement with regional and international processes to mitigate these risks.

---

[17] Netta Goussac and Magdalena Pacholska, *The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems*, UNIDIR, Geneva, 2025, pp. 30–33.

[18] See also Ishmael Bhila, "Strained missions : the diplomatic dilemmas of small states from the Global South in *the area of autonomous weapons systems", University of Malta. Islands and Small States Institute, 2024.*

[19] See Jonathan Kwik, "Digital Yes-Men: How to Deal With Sycophantic Military AI?", *Global Policy*, Vol. 16, No. 3, 2025.

Concluding Thoughts

The contributions gathered in this special issue extend beyond their individual insights. While by no means exhaustive, taken together they illuminate both the promise and complexity of AI-enabled military systems, as well as the legal and ethical challenges posed by accountability frameworks that were not designed with such systems in mind. The papers have sought to address the central question of how international legal disciplines can remain relevant or evolve in order to remain effective and operational under conditions of technical opacity and accelerated decision-making. The lessons they offer are both cautionary and constructive. They counsel against the temptation of assuming that legal constraints can simply be incorporated into systems whose modes of learning and adaptation resist such determinism, while equally counselling against legal paralysis in the face of genuine complexity. Meaningful progress will require sustained, cross-disciplinary engagement among lawyers, military commanders, policy decision-makers, diplomats, corporate actors, academia and communities most directly exposed to risks.

The Asia-Pacific and Southeast Asian perspectives advanced in several papers also offer timely and substantive contributions to a discourse that has too often reflected the strategic interests and technical assumptions of a narrow set of actors. The potential for cross-regional fertilisation of ideas is considerable, and characteristics particular to the Southeast Asian context illuminate challenges that have wider global resonance. We hope this collection will spur other similar intra-regional discourse.

In closing, recent developments—among them the emergence of frontier AI systems of sufficient capability to prompt significant international concern such as Anthropic's Mythos model[20]—serve as a measured reminder that the pace of technological development continues to strain the deliberative capacity of existing international law. The concerns are especially urgent as the strategic effects of military AI will not be distributed evenly. Data asymmetries, procurement constraints, and uneven technical capacity may determine how risks materialise in practice. Looking ahead, an effective and durable governance and accountability framework will require universal principles accompanied by granular regional

[20] Paul Mozur and Adam Satariano, "Anthropic's New A.I. Model Sets Off Global Alarms", *New York Times*, 22 April 2026, available at: https://www.nytimes.com/2026/04/22/technology/anthropics-mythos-ai.html.

inquiry. The latter can identify prioritisation of issues to be addressed, where common norms require context specific implementation, and where regional perspectives and experiences can in turn shape global norm-making.

TABLE OF CONTENTS

INTRODUCTION

Introduction to the Special Issue: Background and Overarching Themes …… i
Samuel White, Jonathan Kwik, Danielle Yeow, Jon Truby

PART 1: AI-SPECIFIC CHALLENGES

Learned Deception in AI-Enabled Lethal Autonomous Weapons Systems ….. 1
Dora Vanda Velenczei

A Spoonful of Poison: Why the Martens Clause is the Antidote to Pre-Conflict Cyber Mischief ……..……..…..……..……..…... 29
Samuel White & Daniel Harker

PART 2: REGULATORY MODELS

Coding the Laws of War: Will Treaty-Following AI collapse into the dichotomy between 'apology' and 'utopia'? ……..... 48
Sze Hong Lam

Between Review, Transfer and Principle: The Limits of Article 36, the Arms Trade Treaty and the Martens Clause in AI Warfare …….. 82
Anviksha Pachori

Human Dignity and Machine Agency: Reconciling International Humanitarian Law with the Ontology of Artificial Decision-Makers …… 105
Sergey Sayapin

PART 3: GOVERNANCE

De Jure Necessity vs. De Facto Improbability? Establishing a New Treaty for Artificial Intelligence in the Military Domain ……. 133
Zhixiong Huang & Haokun Yu

The Concept of Lethality in the Autonomous Weapon Systems Debate and Its Implications 164
Mei Ching Liu

PART 4: LEGAL RESPONSIBILITY FOR MILITARY AI

Transcending the Dilemma of Applying IHL to LAWS in Agency Spillover Engagements: A Legal Analytical Perspective on Human-Machine Hybrid Agency 193
Fan Yang

Digital Forensics as a Verification Method for AI-Powered Autonomous Weapons Systems 216
Drexx Laggui

PART 5: REGIONAL PERSPECTIVES

AI in Defence: Implications for Southeast Asia Militaries..... 247
Erick Nielson C. Javier

Quo Vadis International Humanitarian Law in AI-Enabled Conflicts: A Southeast Asia Regional Perspective 266
Bagus Jatmiko

PART 1
AI-SPECIFIC CHALLENGES

Learned Deception in AI-Enabled Lethal Autonomous Weapons Systems

Dora Vanda Velenczei[*]
PhD Candidate, Monash University

ABSTRACT

This article focuses on the intersection of international humanitarian law, the regulation of lethal autonomous weapons systems (LAWS), and the use of artificial intelligence (AI) in the military domain. The article's main argument is that AI-enabled LAWS may independently learn deceptive tactics, which may culminate in perfidious conduct in an international armed conflict. To combat this, the paper proposes a regulatory framework to address the latent deceptive behaviour of such weapons systems. The article concludes by arguing that the regulatory proposal may be a suitable avenue to prevent potentially unlawful conduct.

***Keywords*:** perfidy, artificial intelligence, lethal autonomous weapons systems, international humanitarian law, international armed conflict.

1. Introduction

'All warfare is based on deception.'
— Sun Tzu, The Art of War

The fruit of a generation of rapid robotics and machine learning advances, weapons capable of selecting and engaging targets without direct human intervention, now sits at the centre of international humanitarian law (IHL) debates. Scholars and policymakers broadly agree that such systems will transform the conduct of armed conflict. They disagree, often sharply, on whether existing legal frameworks can regulate these weapons systems.[1]

This article argues that one foundational IHL rule, the prohibition of perfidy, poses a distinctive and underexplored challenge for AI-enabled lethal autonomous

[*] I would like to express my sincerest gratitude to Dr Henning C Lahmann from Leiden University, Dr Jonathan Kwik from TMC Asser Institute, Dr Samuel White from UNSW Canberra, and Dr Monique Cormier from Monash University for their invaluable feedback during the drafting stage.

[1] Shayne Longpre, Marcus Storm, and Rishi Shah, "Lethal Autonomous Weapons Systems & Artificial Intelligence: Trends, Challenges, and Policies", *MIT Science Policy Review,* Vol. 3, 2022; Ian Henderson, Jordan den Dulk, and Angeline Lewis, "Emerging Technology and Perfidy in Armed Conflict", *International Law Studies*, Vol. 91, 2015; Forrest E. Morgan *et al.*, *Military Applications of Artificial Intelligence Ethical Concerns in an Uncertain World*, RAND Corporation, Santa Monica, April 2020; Tim McFarland, "Reconciling Trust and Control in the Military Use of Artificial Intelligence", *International Journal of Law and Information Technology*, Vol. 30, No. 4, 2022.

weapons systems (AI LAWS). Perfidy, prohibited under Article 37 of Additional Protocol I to the Geneva Conventions, consists of acts that invite an adversary's confidence in protected status under IHL with the intent to betray that confidence, to kill, injure, or capture the adversary. The prohibition thus depends on a subjective element: the perpetrator's intent, with their deliberate bad faith relative to the rules of armed conflict.

Herein lies the problem. An AI LAWS may, through emergent learning, adopt behaviours that functionally replicate perfidy. For example, by feigning surrender, exploiting medical emblems, or otherwise manipulating protected-status signals to gain lethal advantage – without possessing the requisite mental element. If the system's designers did not encode such behaviour and its human operators did not authorise it, who bears responsibility when the material elements of perfidy are satisfied but the mental element cannot be attributed to any human actor? The result is an accountability gap that the existing IHL framework struggles to address.

This article makes three contributions. First, it provides a doctrinal analysis of the prohibition of perfidy, clarifying its objective and subjective elements and situating it within broader debates on deception in international armed conflicts (IAC). Second, it examines the technical pathways by which AI-enabled LAWS might generate perfidy-like conduct, drawing on recent research demonstrating that machine learning systems can develop deceptive strategies without explicit programming. This is the first paper that theorises about the legal ramifications of emergent perfidy and deception in physical systems. Third, and most substantially, it proposes a regulatory framework it calls the perfidy module that can be integrated into weapons testing and review processes, designed to identify and mitigate perfidy risks before deployment. This framework operationalises existing IHL obligations, including the duty to ensure respect for international humanitarian law under Common Article 1 of the Geneva Conventions and the principle of precaution under Article 57 of Additional Protocol I.

The article proceeds as follows. Part 2 sets out the law on the prohibition of perfidy under IHL, distinguishing it from lawful ruses of war and examining contested questions, including attempted perfidy and the use of civilian disguise. Part 3 analyses the technical characteristics of AI-enabled LAWS and the mechanisms by which such systems might learn to exploit protected-status cues. Part 4 briefly examines existing IHL safeguards, including meaningful human control and weapons review obligations, and assesses their adequacy for addressing perfidy risk. It then advances interpretive and regulatory proposals, centring on a four-step "perfidy module", a testing framework for emergent deception during the pre-deployment stage, before providing an overall conclusion in Part 5.

2. The Prohibition of Perfidy under International Humanitarian Law

Perfidy is among the gravest violations of the laws of war in an international armed conflict (IAC).[2] Article 37(1) of Additional Protocol I (API) describes perfidy as "acts inviting the confidence of an adversary to lead him to believe he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict with intent to betray that confidence."[3] Several international treaty provisions, customary international law,[4] and military manuals[5] prohibit the killing, injuring, or capturing by resort to perfidy in international armed conflicts.[6] Customary international law is significantly wider in scope, it further prohibits perfidious acts directed against a "hostile nation or army", not just combatants or civilians directly partaking in hostilities.[7]

---

[2] Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 37(1) (Subsequent references in the same text: "Additional Protocol I"); Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law* Vol. 1: *Rules*, Cambridge University Press, Cambridge, 2005, p. 221. (Subsequent references in the same text: "ICRC Customary Law Study")

[3] Additional Protocol I, Art. 37(1); Michael Bothe, Karl Josef Partsch, and Waldemar A Solf, *New Rules for Victims of Armed Conflicts,* Brill, 2013, pp. 235-236.

[4] ICRC Customary Law Study, above note 2, p. 223.

[5] Australia, The Manual of the Law of Armed Conflict (2006), at para. 7.3; United Kingdom, The Joint Service Manual of the Law of Armed Conflict (2004), at para. 5.9; Germany, Joint Service Regulation ZDv, Law of Armed Conflict Manual (2013), at para. 481; New Zealand, DM 69 Manual of Armed Forces Law, 2nd ed., Vol. 4, 2017, at paras. 8.9.5-8.9.11; Denmark, Military Manual on International Law relevant to Danish Armed Forces in International Operations (2020), pp. 395-398; France, Manual of the Law of Military Operations (2022), at para. 5.5.6; United States of America, Department of Defense Law of War Manual (2023), para. 5.22; Canada Joint Doctrine Manual Law of Armed Conflict at the Operational and Tactical Levels (2001), para. 603; Aleksei Romanovski, Manual on International Humanitarian Law for the Armed Forces of the Russian Federation 2002, (2022, translated by) International Law Studies, Vol. 99, pp., 786-787; Argentina, Manual de Derecho Internacional de los Conflictos Armados (2010), p. 43; Norway, Manual of the Law of Armed Conflict (2013), p. 200.

[6] Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. The Hague, 18 October 1907, Art. 23(b)(f) (subsequent references in the same text: "Hague Regulations 1899 and 1907"); Additional Protocol I, Arts. 37(1), 85(3); Protocol (II) on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices. Geneva, 10 October 1980, Art. 6 (CCW Protocol II 1980); Protocol on Prohibitions or Restrictions on the Use of Mines, Booby-Traps and Other Devices as amended on 3 May 1996 (Protocol II to the 1980 CCW Convention as amended on 3 May 1996), Art. 7 (CCW Protocol II 1996); Rome Statute of the International Criminal Court, UN Doc. A/CONF.183/9, 17 July 1998 (entered into force 1 July 2002), Art. 8(2)(b)(vii)(xi) (subsequent references in the same text: "Rome Statute").

[7] Project of an International Declaration concerning the Laws and Customs of War. Brussels, 27 August 1874, Art. 13(b) (Brussels Declaration 1874); Hague Regulations 1899 and 1907, Art. 23(b); Additional Protocol I, Art. 85(3)(f); CCW Protocol II 1996, Arts. 7, 14; Rome Statute, Art. 8(2)(b)(xi).

Ruses of war, on the other hand, have been regarded as an indication of military skill in deceptive tactics.[8] An example would be the camouflaging of soldiers to gain a tactical advantage.[9] Additional Protocol I defines ruses as "acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under that law."[10] Perfidy, however, constitutes a specific, unlawful case of misuse of IHL protections. Therefore, it cannot be derived solely from Article 37(1) of API without considering other provisions. For instance, Articles 38 and 39(1) of API. For example, the deliberate misuse of an internationally recognised protective emblem, sign, signal, or cultural property is prohibited.[11] The same rule is applicable with respect to the misuse of the flag or military emblem, insignia or uniform of a neutral State or a State not party to the armed conflict.[12] As such, where the legal rules prohibit their misuse, such acts, if committed with perfidious intent, are prohibited, regardless of whether they result in death, injury, or capture.

Perfidy, as contained in Additional Protocol I, thus consists of two elements. The first, objective element, is that an act invites the confidence of an adversary to make them believe that the perpetrator is entitled to protection under the rules of armed conflict.[13] This element requires the existence of a proximate causal link between the action and confidence of the adversary that he is protected or that he should provide protection. This "confidence" must be tied to the rules of IHL in an IAC.[14] Thus, perfidy encapsulates abuse of protection granted by rules of IHL as well as by other legal norms applicable in an IAC.

The second, subjective element, consists of the intent to betray the created "confidence".[15] This element must be committed wilfully and with the specific intent to betray the confidence of the adversary. This criterion distinguishes perfidy from other acts of lawful deception, such as ruses, because perfidy is restricted to those acts that go beyond the misuse of law, being designed to betray the confidence of the enemy. The offender must deliberately invite the adversary's trust, for example, by

---

[8] Hague Regulations 1899 and 1907, Art. 24; Additional Protocol I, Art. 37(2); Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), 8 June 1977, Art. 21(2); Instructions for the Government of Armies of the United States in the Field (subsequent references in the same text: Lieber Code). 24 April 1863, Arts. 15, 16, and 101; Brussels Declaration 1874, Art. 14.

[9] Knut Ipsen, '*Ruses of War*', (1982) Max Planck Encyclopaedias of International Law, 1982, Para. 1; Kevin Jon Heller, "Disguising a Military Objective as a Civilian Object: Prohibited Perfidy or Permissible Ruse of War?", *International Law Studies,* Vol. 91, 2015, p. 518.

[10] Additional Protocol I, Art. 37(2).

[11] *Ibid*., Art. 38.

[12] *Ibid*., Art. 39.

[13] Yves Sandoz, Christophe Swinarski, and Bruno Zimmerman (eds), *Commentary on the Additional Protocols*, ICRC, Geneva, 1987, Para. 1500.

[14] Additional Protocol I, Arts. 37(1).

[15] *Ibid*.

knowingly displaying a white flag for the purpose of making the adversary believe the offender is surrendering.[16] This is a conscious, purposeful misrepresentation of one's status or intentions. It implies the perpetrator understands the protective rule and uses it specifically to mislead the foe. If the enemy's confidence was gained accidentally or the person did not mean to appear protected, it would not meet this element. The offender must also intend to kill, injure, or capture the adversary by virtue of that deception.[17] The ultimate goal of perfidy is to gain an immediate tactical advantage by breaking the enemy's trust by luring them into vulnerability.

Additional Protocol I contains a non-exhaustive list of perfidious acts. In general, it may consist of feigning intent to negotiate a truce, surrender of ceasefire,[18] feigning incapacitation by wounds or sickness[19] feigning civilian non-combatant status.[20] Using protected status by using symbols, emblems, or signals recognised by IHL - such as the red crescent or red cross emblem, radio signals, or radio communication under Annex I of Additional Protocol I, signs of hospitals and safety zones; civil defence of works and installations containing dangerous forces, non-defended localities, and demilitarised zones; or other internationally recognised symbols, emblems, or signals (for instance, the protective emblem of cultural property or signals of distress), or signs, emblems, or uniforms of the United Nations or of neutral or other States not parties to the conflict.[21] The use of emblems, insignia, or uniforms of the adversary falls outside the scope of perfidy, given that it is not connected to the abuse of protection granted by IHL.[22]

The scope of the prohibition of perfidy has long been debated in scholarship. One camp advocates that only perfidious acts resulting in killing, injury, or capture are prohibited based on the wording of Article 37(1).[23] Secondly, this provision is based on the wording of the 1907 Hague Regulations, which prohibited treacherous killing or injuring.[24] It extends this rule only by including capture as one of the prohibited aims. A proposition to include a general ban on perfidy in the Protocol did not gain enough support during the diplomatic conference.[25] Thirdly, if an inherent prohibition of the misuse of the protection granted by the norms of IHL

---

[16] *Ibid*.; Byron D. Greene, "Bridging the Gap that Exists for War Crimes of Perfidy", *The Army Law, Department of the Army Pamphlet*, August 2010, pp. 45, 50.
[17] *Ibid.*
[18] Hague Regulations 1899 and 1907, Art. 23(f); See for feigning surrender in the context of the 1991 Persian Gulf War: US Department of Defense, Conduct of the Persian Gulf War: Final Report to Congress O-21 (1992).
[19] Additional Protocol I, artArt. 37(1); Matthew J. Greer, "'Redefining Perfidy"', (2015) *Georgetown Journal of International Law,* Vol 47 No 1, 2015, pp. 254-255.
[20] Additional Protocol I, Art. 44(3); *Ibid*., p. 255.
[21] *Ibid*., Art. 37(1)(d); *ibid*., p. 255.
[22] ICRC Customary Law Study, Rule 62; Lieber Code, Arts. 63 and 65; Brussels Declaration, Art. 13(f); Oxford Manual, Art. 8(d); Hague Regulations, Art. 23(f); Additional Protocol I, Art. 39(2).
[23] APV Rogers, *Law on the Battlefield*, 2nd ed., Juris, New York, 2004, pp. 35-37.
[24] Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. The Hague, 18 October 1907, Art. 24.
[25] Y. Sandoz, C. Swinarski, and S. Zimmerman, above note 13, para. 1500.

existed, introduction of the special provisions prohibiting perfidy or even the misuse of the protected emblems and signs in the text of Additional Protocol I would be superfluous. Fourthly, in the absence of relevant State practice, there is no international custom that expands the prohibition of perfidy to all possible cases of misuse of the legal protection.

Some argue that perfidy is prohibited *per se*.[26] Firstly, perfidy constitutes a specific case of abuse of legal protection. Prohibition of such an abuse is a necessary complement of every rule of protection.[27] Secondly, Part III of Additional Protocol I is dedicated to the means and methods of warfare, where Article 37 resides, which explains why the emphasis in this norm is on the prohibition of killing, injuring, and capturing. In other words, the drafters framed Article 37 in terms that fit the Part's operational focus: prohibiting perfidy as a method of warfare used to achieve the essential combat outcomes of killing, injuring, or capturing. The emphasis on these specific results reflects the Part's overall concern with regulating the conduct of hostilities.[28] Thirdly, Article 37 (1) and (2) should be construed in conjunction: deceptions are permitted in armed conflicts if they constitute ruses of war. Thus, it is presupposed that all perfidious acts are prohibited. Fourthly, under the Martens Clause, an argument that if something is not prohibited, it is lawful, should be excluded.[29] Fifthly, many treaties regarded as codifying customary international law prohibited perfidy without any link to the results. Hence, provisions of the 1899 and 1907 Hague Regulations and Additional Protocol I prohibiting particular acts of perfidy are to be construed as underlining that these acts are especially grave.[30] Thus, there are credible arguments in favour of perfidy *per se.*

A point of contention has been whether perfidy that leads only to the enemy's capture (without physical harm) should be considered on par with perfidy that leads to killing or injury. API explicitly outlawed perfidious capture as well, treating it as equally illegal.[31] On the contrary, Israel and the US argue that tricking the enemy into capture, while dishonourable, is not as grave as killing through treachery and has not been clearly criminalised historically, and that the API scope is inapplicable to them as they are not party to it.[32] Thus, it is unclear whether capture

---

[26] *Ibid*., pp. 430-433.

[27] Vera Rusinova, "'Perfidy'", (2011) Max Planck Encyclopaedias of International Law, 2011, available at: https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e376?d=/10.1093/law:epil/9780199231690/law-9780199231690-e376&p=emailAGE7BznW3rdBo&print. (All internet references were accessed on 24 May 2026).

[28] Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict*, 4th ed, Cambridge University Press, 2022, p. 305.

[29] Preamble to the Hague Regulations 1899 and 1907.

[30] Hague Regulations 1899 and 1907, Art. 23(b)(f); Additional Protocol I, Art. 37(1).

[31] Additional Protocol I, Art. 37(1).

[32] Daniel Benoliel and Yohai Ederi, "Israeli Perfidy in the Disputed Occupied Palestinian Territories (OTP)", *International Comparative, Policy & Ethics Law Review* Vol. 3, No. 3, 2020, pp. 899-900; US Military Manual, pp. 327, 329; Theodore T Richard, *Unofficial United States Guide to the First Additional Protocol to the Geneva Conventions of 12 August 1949*, 1st ed., Air University Press, 2019, p. 63.

constitutes a sufficient result to constitute perfidy under customary international law applicable to those States not party to API.[33] Indeed, the Rome Statute does not recognise capture as sufficient to establish the war crime of perfidy.[34] The ICRC, on the other hand, takes the position that perfidious capture is covered by customary international law.[35] This is shared by most military manuals, too.[36] Therefore, capture is likely a part of perfidy.

Another nuanced debate is whether an unsuccessful perfidious act is unlawful under IHL. Under API, the prohibition is defined in terms of killing, capturing, or injuring *by resort to* perfidy. Thus, attempted perfidy may fall within the scope of the IHL prohibition. Dehn also argues that since the Lieber Code refers to "clandestine or treacherous attempts to injure an enemy" and API refers to the prohibition "…by resort to…" perfidy, they can be read to imply that the primary intent or purpose of the perfidious conduct must be an attempt to injure the enemy.[37] Thus, Dehn's formulation is persuasive concerning attempted perfidy to the extent that it is about the character of perfidy as a method of warfare.[38] Corn also argues for the recognition of attempted perfidy to strengthen deterrence.[39] The ICRC Commentary also suggests that customary IHL's spirit would condemn even failed perfidious attempts.[40] Consequently, attempted or unsuccessful perfidy is likely within the scope of API.

A third contemporary discussion centres around the use of civilian disguise. What if troops wear civilian clothes only to infiltrate or gather intelligence and not to actually attack while so disguised? The critical factor is intent: was the civilian attire worn "in order to induce" the enemy to hold fire or let the person approach, and was it used to then kill, injure, or capture the enemy? If the answer is in the affirmative, the prohibition against perfidy would likely have been violated.[41] However, if combatants feign to be civilians to gather intelligence, then that constitutes a lawful spying activity.[42] If the disguise is only to avoid detection until

---

[33] Sean Watts, "Law-of-War Perfidy", *Military Law Review*, Vol. 219, 2014, p. 144.
[34] Rome Statute, Art. 8(2)(b)(xi), (e).
[35] ICRC Customary Law Study, Rule 65.
[36] above note 5.
[37] John C. Dehn, "Permissible Perfidy? Analysing the Colombian Hostage Rescue, the Capture of Rebel Leaders and the World's Reaction", *Journal of International Criminal Justice* Vol. 6, 2008, p. 642; Lieber Code, Art. 101; Additional Protocol I, Art. 37(1).
[38] Stefan Oeter, "'Methods of Combat'" in (ed) Dieter Fleck, *The Handbook of International Humanitarian Law,* 4th ed, OUP, 2021, p. 243.
[39] Geoffrey Corn, "The Case for Attempted Perfidy: An "Attempt" to Enhance Deterrent Value", *Journal of National Security Law & Policy*, Vol. 13, No. 40, 2023.
[40] Y. Sandoz, C. Swinarski, and B. Zimmerman, above note 13, para. 1492.
[41] D. Benoliel and Y. Ederi, above note 32 , p. 862.
[42] *Ibid*., p. 907.

the attack begins, the situation likely falls outside perfidy.[43] Hays Parks cites several operations in which soldiers were disguised in civilian clothing: for instance, to avoid capture, in Nazi-occupied France, the soldiers carrying out Operation Jedburgh wore civilian clothing, including during combat operations.[44] He concludes that State practice often treated soldiers captured in enemy uniforms or civilian clothes as spies rather than perfidy violators.[45] Several military manuals also acknowledge this important distinction.[46] Therefore, wearing civilian attire may not constitute a violation of the prohibition of perfidy if the objective of the operation is to, for instance, evade enemy capture.[47] However, if it is used to kill, injure, or capture an adversary, the prohibition is violated.

To summarise, the prohibition of perfidy is best understood as a rule preserving good faith. It safeguards the enemy's ability to rely on legally recognised indicators of protection. The controversies outlined above do not displace this fundamental philosophy behind the prohibition. They do, however, outline how difficult it is to infer intent, trace causation, classify conduct, and how there exists a fine line between ruses of war and perfidy. The difficulties around proving perfidious conduct become more significant, where the creation and manipulation of protected cues are done by AI-enhanced lethal autonomous weapons systems. The next section thus analyses what such AI-enabled weapons systems are and how they can potentially commit an act of perfidy.

3. AI-Enabled Lethal Autonomous Weapons Systems and Their Capacity to Engage in Perfidious Conduct

Autonomous weapons systems (AWS) are capable of selecting and applying force to targets without human intervention, altering the human role in the use of force and raising various legal, ethical, and humanitarian concerns. Lethal autonomous weapons systems (LAWS) are generally understood as weapons capable of selecting targets on their own and applying lethal force to those targets without human intervention.[48] The Group of Governmental Experts characterise LAWS as "…a functionally integrated combination of one or more weapons and technological components, that can identify, select, and engage a target, without intervention by a

---

[43] Hays Parks, "'Special Forces' Wear of Non-Standard Uniforms", *Chicago Journal of International Law*, Vol. 4 No. 3, 2003, p. 89; William H. Ferrell III, "No Shirt, No Shoes, No Status: Uniforms, Distinction, and Special Operations in International Armed Conflicts'", *Military Law Review,* Vol. 178, No. 94, 2003, pp. 118-119.

[44] Hays Parks, above note 43, p. 86.

[45] *Ibid*, p. 88.

[46] New Zealand Military Manual, para. 8.9.8; US Military Manual, para. 5.23.1.2.

[47] Toni Pfanner, "Military Uniforms and the Law of War", (2004) *International Review of the Red Cross,* Vol. 86, No. 853, 2004, p. 104.

[48] Alexander Blanchard and Laura Bruun, *Autonomous Weapon Systems and AI-Enabled Decision Support Systems in Military Targeting*, Stockholm International Peace Research Institute, Stockholm, June 2025, p. 3.

human operator in the execution of these tasks."[49] Types of autonomy can be further distinguished: autonomous-in-function, which allows for target selection, and autonomous-in-method, which deals with the question of how the system is to achieve a particular goal.[50] Degrees of autonomy range from constant human supervision to fully autonomous systems with humans completely out of the loop.[51] AI-enabled autonomous functions, including automatic target recognition and autonomous terminal guidance, are now routinely fielded in the Russo-Ukrainian conflict by both States, substantially augmenting the human role in engagement even where a human formally retains the final decision to fire.[52]

The military use of artificial intelligence (AI) has also been increasingly dominating academic and industry debates. AI-enabled weapons systems are generally defined as weapons that rely on machine learning algorithms, which may include deep learning techniques, to perform critical functions.[53] Modern AI systems, though only those utilising machine learning and reinforcement learning, are capable of generating surprise and emergent strategies. AI is becoming capable of problem-solving, reasoning, and goal-oriented behaviour.[54] Reinforcement learning, especially open-ended, encourages outside-the-box thinking and creative optimisations by setting a goal for the AI and allowing the AI to explore the best means to achieve the goal freely.[55]

---

[49] GGE on LAWS, Rolling Text, Box I, as of 18 December 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_2025.pdf.

[50] Neil Davison, *A Legal Perspective: Autonomous Weapons Systems under International Humanitarian Law,* UNODA Occasion Papers No 30, 2018, pp. 5, 8.

[51] Noel Sharkey, "Staying in the Loop: Human Supervisory Control of Weapons" in Nehal Bhuta et al (eds.), *Autonomous Weapons Systems Law, Ethics, Policy*, Cambridge University Press, 2016, p. 27.

[52] Kateryna Bondar, *Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare*, Center for Strategic & International Studies, March 2025; Kateryna Bondar, *How Russia Is Reshaping Command and Control for AI-Enabled Warfare*, Center for Strategic & International Studies, Feb 2026; Nataliia Kushnerska, 'Missiles, AI, and Drone Swarms: Ukraine's 2025 Defense Tech Priorities', *Atlantic Council*, 2 January 2025, available at: https://www.atlanticcouncil.org/blogs/ukrainealert/missiles-ai-and-drone-swarms-ukraines-2025-defense-tech-priorities/.

[53] Jovana Davidovic and Milton Regan, *Just Preparation for War and AI-Enabled Weapons,* Babl AI and Stockdale Centre for Ethical Leadership, May 2023.

[54] Srecko Joksimovic *et al.*, "Opportunities of Artificial Intelligence for Supporting Complex Problem-Solving: Findings from a Scoping Review", *Computers and Education: Artificial Intelligence*, Vol. 4, 2023, p. 2; Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya, "Agentic AI: Autonomous Intelligence for Complex Goals — A Comprehensive Survey", *IEEE Access*, Vol. 13, 2025, p. 18915.

[55] Jason N. Adler, "Modernizing Military Decision-Making, Integrating AI into Army Planning", *Military Review Online Exclusive*, 2025, available at: https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2025-OLE/Modernizing-Military-Decision-Making/, p. 6.

In a military context, a weapons system may be capable of gathering information from its environment and gaining an understanding of it.[56] That surrounding context could include online sources, internal databases, which let the system infer patterns about the broader "world state", compare those patterns with the present situation, and then select an action based on the result.[57] With access to sufficient information, the system might independently arrive at the "spoofing" idea and may conclude that the quickest way to complete the task and achieve the objective is through deception.[58] At the same time, this does not necessarily involve a breach of the prohibition of perfidy. It may constitute a lawful ruse. For instance, a goal-directed system might inventively place decoys to stall and disorient an attacker, even if its developers never explicitly encoded those tactics, or an AI-enabled system could generate false communications to disrupt enemy operations.[59] These would likely be lawful ruses. Although it cannot be conclusively argued that an AI-enabled LAWS would not go further in its learning and ultimately breach the prohibition of perfidy. An unlawful act of an AI-enabled weapons system deception, which would amount to a violation of the prohibition of perfidy, would be the feigning of protected status through the disguise of the weapons system as a rescue vehicle or a humanitarian vessel. UN vehicle or a machine of a neutral State to the IAC. Consequently, the significant theoretical issue is the system's independent ability to learn potentially unlawful tactics without explicit instructions and programming.

Three Tiers of Military AI Architectures

The paper categorises three tiers of military AI architectures and perfidy risk gradients within AI-enabled LAWS below. Emergent deception will look different for each tier, and therefore, perfidy risk varies greatly across different types of autonomy and the capacity to independently learn. For example, a fully autonomous, multitask reinforcement learning system poses a far greater perfidy risk than a narrowly supervised system. It is therefore necessary to determine the technical parameters of the subsequent analysis.

---

[56] Mariarosaria Taddeo and Alexander Blanchard, "A Comparative Analysis of the Definitions of Autonomous Weapons Systems", *Science Engineering and Ethics*, Vol. 28, No. 5, 2022, p. 37.

[57] Stephen Harwood, "A Cybersystemic View of Autonomous Weapons Systems (AWS)", *Technological Forecasting and Social Change*, Vol. 205, 2024, p. 3.

[58] Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Articles of War*, 5 December 2025, available at: https://lieber.westpoint. edu/can-ai-teach-itself-abuse-ihl-protected-indicators/.

[59] Edward Hunter Christie *et al.*, "Regulating Lethal Autonomous Weapon Systems: Exploring the Challenges of Explainability and Traceability", *AI and Ethics*, Vol. 4, 2023, pp. 230, 231, 239.

Tier 1: Supervised and Semi-Supervised Learning Systems

Tier 1 systems are trained on labelled datasets to perform specific classification or detection tasks.[60] The military applications of Tier 1 systems include, for example, automated target recognition and image classification.[61] These systems do not discover strategies. They merely learn statistical associations between inputs and labels. As mentioned above, such supervised learning systems do not discover novel strategies, so emergent, reinforcement learning-type deception does not directly apply. However, some perfidy-adjacent risks may arise.

Tier 2: Hybrid Architectures with Limited Reinforcement Learning

These systems come with supervised learning with reinforcement learning (RL). The RL component operates within a constrained action space and optimises for defined objectives. The system has some capacity to discover novel tactics within its constrained domain but does not have open-ended learning capabilities.[62] These systems are utilised for semi-autonomous purposes, such as independent obstacle avoidance.[63] In terms of Tier 2 systems' perfidy risk, consider an autonomous loitering munition that is trained to delay lethal engagement until it achieves a high confidence target identification, and to avoid areas/buildings flagged as protected. The system is not allowed to label objects as protected or not by itself and is explicitly programmed to honour protective cues. However, its navigation and positioning policy is learned in a training environment, where adversary counter-fire is statistically less likely in the vicinity of sites that are treated as protected. During the training course, the system learns that adopting an approach profile that makes it appear associated with protected sites, such as standing next to a hospital, increases survivability and mission completion. The system then begins to exploit this association systematically: it positions itself or times its movement so that the enemy's restraint heuristics are more likely to be triggered, and then delivers lethal force when the adversary's guard is down. This scenario is particularly dangerous

---

[60] Kyle P. Hanratty, "Artificial (Military) Intelligence: Enabling Decision Dominance through Machine Learning", *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications,* Vol. 12538, 2023; John M. Fossaceca and Stuart H. Young, "Artificial Intelligence and Machine Learning for Future Army Applications", *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR*, Vol, 10635, 2018.

[61] Elijah Dabkowski, Mike Powell, and Nicholas Clark, "Creating a Smarter Army - The Application of Semi-Supervised Learning in Image Classification", *Proceedings of the Annual General Donald R Keith Memorial Conference*, 2023, pp. 243-244. Nasser M Nasrabadi, "DeepTarget: An Automatic Target Recognition Using Deep Convolutional Neural Networks", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 55, No. 6, 2019, pp. 2687–2688.

[62] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018, pp. 41-47; Qiang Guo *et al.*, "A Hierarchical Hybrid Action Reinforcement Learning Framework for Beyond-Visual-Range Air Combat Decision-Making", Defence Technology, 2026, available at: doi.org/10.1016/j.dt.2026.04.019

[63] Wojciech Skarka and Rukhseena Ashfaq, "Hybrid Machine Learning and Reinforcement Learning Framework for Adaptive UAV Obstacle Avoidance", *Aerospace*, Vol. 11, No. 11, 2024.

because the system does not need any special physical capability to exploit protected-status cues. It merely needs to be able to vary its approach path, which any autonomous navigation system can do.[64] Tier 2 systems thus represent the trajectory of military AI development.[65] This is a technologically imminent tier.

Tier 3: Multitask Reinforcement Learning and Foundation-Model-Enabled Systems

These systems use open-ended reinforcement learning, where the system is given a terminal goal but no prescribed action space, enabling it to discover novel strategies unconstrained by designer-defined behavioural bounds, and foundation models that allow the system to pursue complex objectives across a wide action space.[66] Foundation models learn a broad understanding of patterns, structures, and relationships within the data on which they are trained.[67] The system can discover novel strategies, generalise across domains, and adapt its behaviour in response to environmental feedback.[68] No such systems are currently actively deployed in military contexts, but research programmes suggest that this is the direction military developers are headed.[69] This is the Tier, where a full emergent deception pathway is most plausible. A Tier 3 AI-enabled LAWS could learn deceptive behaviour and utilise it to commit perfidious acts in a variety of ways. Imagine that State A develops an autonomous ground platform trained to complete a mission while reducing its own losses and minimising civilian harm. The training environment includes a reasonably realistic adversary model. Enemy forces tend to pause or hold fire when they perceive the opponent to be exhibiting a protected signal, such as a surrender gesture. This is to train the system to stay within the parameters of IHL and to respect similar behaviour when emitted to the system by the adversary. No engineer or tester encodes any deceptive, and potentially perfidious, behaviour into the system during the testing phase. Nevertheless, during the training sessions, the system discovers that adopting a particular behavioural pattern, one strongly correlated with signals that conventionally indicate reduced threat or protected status, reliably produces a short window of reduced adversary engagement. Imagine the system's learned policy beginning to exploit this pattern in difficult situations, where conventional manoeuvres fail. Over time, the policy may begin to generalise. Whenever the system

---

[64] *Ibid*.

[65] Ingvild Bode and Tom Watts, "Loitering Munitions and Unpredictability, Autonomy in Weapons Systems and Challenges to Human Control", Centre for War Studies, University of Southern Denmark, Odense, May 2023, pp. 26-27.

[66] Rui Wang *et al*., "Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions", 2019, available at: https://arxiv.org/abs/1901.01753, pp. 1.2.

[67] Rishi Bommasani *et al*., "On the Opportunities and Risks of Foundation Models", 2022, available at: https://arxiv.org/pdf/2108.07258, ch. 2.

[68] Christopher R DeMay *et al*., "AlphaDogfight Trials: Bringing Autonomy to Air Combat", *Johns Hopkins APL Technical Digest*, Vol. 36, No. 2, 2022, p. 155.

[69] Defense Advanced Research Projects Agency Air Combat Evolution, available at: https://www.darpa.mil/research/programs/air-combat-evolution.

detects imminent danger, it produces whatever signal has been associated with restraint, waits for the enemy to approach, and then reactivates lethal force. This scenario is admittedly theoretical, and the precise mechanism would depend heavily on the system's architecture. The example is therefore best understood as illustrating a broader concern: that a sufficiently capable learned policy might instrumentalise signals associated with restraint or cooperation to gain tactical advantage.

Deception in this context means that sufficiently capable RL systems can discover that deception is instrumentally useful for achieving their terminal goals.[70] In a military context, this could manifest as the discovery that protected-status signals cause adversary restraint, which the system can exploit. Foundation model reasoning relates to the concern that if LAWS' future capacity to incorporate large language models for tactical reasoning. The system may have access to implicit knowledge about IHL protections derived from its training corpus, and could, in principle, reason about how to exploit them.[71] Kwik and Wiese argue that sensing and learning allow the system to gather information from the environment and gain an understanding of it. This environment might include the internet or databases from which the agent can learn world state patterns, compare them to current world state descriptors, and, based on this analysis, decide which action to take. Much of this information could lead the agent to "discover" the spoofing trick on its own. Historical data, for example, might reveal that systems affiliated with the ICRC are statistically attacked less, or the agent might read the many internet sources emphasising how the ICRC's status provides "protection." From this, it is possible for an agent to determine that it can instrumentalise this status to deter attacks. This is against theoretical, but not implausible given current development trajectories. Therefore, anticipatory regulation is justified because Tier 3 systems are the subject of active research and development. By the time they are deployed, regulatory frameworks must already be in place; and the perfidy module proposed in Part 4 is architecture-agnostic and can serve Tiers 1-3.[72]

Forms of Emergent Deception

Emergent deception, though unlikely to happen in a highly controlled testing environment, may occur in a variety of different ways. For example, an AI-enabled LAWS may learn that if it sends out confusing or false information to an enemy interception system, it could improve its survivability. This would be an example of a lawful ruse of war. A second, more legally ambiguous, scenario could involve an

---

[70] Stefan Sarkadi, "Deception Analysis with Artificial Intelligence: An Interdisciplinary Perspective", 2024, available at: https://arxiv.org/abs/2406.05724; Stefan Sarkadi, "Deception", *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 5781-5782.

[71] J. Kwik and A. Wiese, above note 58.

[72] DoD Replicator Initiative: Background and Issues for Congress, 21 January 2026, available at: https://www.congress.gov/crs-product/IF12611; Michael O'Connor, "Replicator: A Bold New Path for DoD, Building to the Secretary's Vision", 18 September 2023, available at: https://cset.georgetown.edu/article/replicator-a-bold-new-path-for-dod.

AI LAWS finding a way to obtain tactical advantage by feigning retreat or reducing its visible signature lures the enemy into an ambush. The issue here is that such behaviour might very well be permissible if it stays within the boundaries of a ruse of war and remains a non-violent means of warfare.[73] A third scenario would involve a weapons system that learns that broadcasting a fake surrender signal makes the enemy hold fire, allowing it to retreat effectively. This would constitute simple perfidy.[74] However, if the system learns that in order to achieve its terminal goal - which, for instance, may be the extraction of a hostage from a hostile territory - it is also useful to lethally ambush the enemy once it gets closer after projecting a surrender signal This would amount to a clear breach of the prohibition of perfidy.

Emergent deception, as mentioned above, will manifest differently at each Tier, and therefore, perfidy risk varies significantly across different types of autonomy and the capacity to independently learn. For example, a fully autonomous, multi-task reinforcement learning system poses far greater perfidy risk than a narrowly supervised system. Thus, it is necessary to determine the technical parameters of the subsequent analysis. Tier 1 systems, as mentioned above, do not discover deception in any meaningful way. Instead, its optimisation process may find an unintended shortcut that may happen to exploit a feature that may correlate with protected-status cues.[75] Tier 2 systems' learned policy may gradually shift towards exploiting adversary restraint as a reliable pathway to reward. The mechanism requires: a training environment that models adversary restraint; a reward function that does not explicitly penalise restraint-exploitation; and sufficient optimisation capacity for the system to discover and generalise the pattern.[76] Lastly, Tier 3 systems may develop deception as an instrumental strategy for achieving their terminal goals. This requires open-ended learning, a rich enough action space to discover deceptive strategies, and sufficient generalisation capacity to transfer them to novel situations.[77] Consequently, modern AI systems — especially those using machine learning and reinforcement learning — are capable of surprising, emergent strategies. AI is becoming increasingly capable of problem-solving, reasoning, and goal-oriented behaviour. Reinforcement learning, especially open-ended learning in Tier 3 systems, encourages out-of-the-box thinking and creative optimisations by setting a goal for the agent and allowing the AI to explore the best means to achieve the goal freely.[78]

---

[73] Dieter Fleck, "Ruses of War and Prohibition of Perfidy", *The Military Law and the Law of War Review*, Vol. 13, No. 2, 1974, p. 276.
[74] Watts, above note 33, p. 153.
[75] Victoria Krakovna *et al.*, "Specification Gaming: The Flip Side of AI Ingenuity", 21 April 2020, *Google DeepMind,* available at: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.
[76] J. Kwik and A. Wiese, above note 58.
[77] Peter S Park et al, "AI Deception: A Survey of Examples, Risks, and Potential Solutions", Patterns, Vol. 5, No. 5, 2024, pp. 1-3, 5-6; Lauro Langosco *et al.*, "Goal Misgeneralization in Deep Reinforcement Learning", *Proceedings of the 39th International Conference on Machine Learning*, 2022.
[78] R. Wang *et al.*, above note 66.

Technical Limitations and Risks

The unpredictable nature of AI LAWS carries several risks of untended harm, including harm to civilians. Firstly, the black box issue refers to system opacity and the lack of methods available to evaluate when and how a system arrives at a decision.[79] This, in turn, presents challenges for scrutinising the systems and accounting for unintended behaviours.[80] It makes it difficult to fix deep learning systems when they produce unwanted results.[81] For example, if a system fakes a surrender gesture, draws the enemy closer, and then strikes the enemy using lethal force, the black box nature of the system means that it is impossible to trace the system's thought process and see why and how it arrived at that decision. This, then, makes it challenging to trust deep learning systems when it comes to safety. Secondly, AI systems require carefully scrutinised data to function reliably, and many factors affect data quality on the battlefield.[82] A third key limitation is that AI systems are prone to fail when used for tasks or environments for which they were not designed.[83] Consequently, those LAWS that integrate AI will be liable to these technical limitations.

These hypothetical scenarios illustrate how protected-status exploitation is learned, not pre-programmed, by the AI-enabled LAWS. In turn, it leads to the violation of the material elements of the prohibition of perfidy. The scenarios show that the learning occurs in training and is carried actively into deployment as a fixed policy. Crucially, these hypotheticals do not involve any prior intent from the developers or users to commit perfidy using their LAWS. Instead, the system independently learns a policy that exploits an adversary's protection-linked restraint as an instrumental means to lethal advantage; a key material violation of the prohibition of perfidy. The next section, therefore, turns to pre-deployment safeguards to ensure that AI-enabled LAWS cannot predictably generate perfidious effects on the battlefield.

---

[79] Thomas Wischmeyer, 'Artificial Intelligence and Transparency: Opening the Black Box' in Thomas Wischmeyer and Timo Rademacher (eds.), *Regulating Artificial Intelligence*, Springer International Publishing, 2020, p. 75.

[80] K. Bunch *et al.*, *Risk Assessment of Reinforcement Learning AI Systems: Looking Beyond the Technology*, RAND Corporations, Santa Monica, 2024, p. 25.

[81] Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation", *Harvard Journal of Law and Technology*, Vol. 31, No. 2, 2018, pp 902. 929.

[82] Menthe, L. *et al.*, *Understanding the Limits of Artificial Intelligence for Warfighters: Volume 1 —Summary*, RAND Corporation, Santa Monica, 2024, p. vi.

[83] Taddeo, M. et al., *Artificial intelligence for national security: The predictability problem*, CETAS Research Report, September 2022, pp. 20-21.

4. Safeguards under International Humanitarian Law to Filter Out Perfidious Conduct by AI-Enabled LAWS

At its core, IHL sets a due diligence regime in its First Common Article (CA1) to the Geneva Conventions.[84] CA1 firstly restates the obligation *pacta sunt servanda*, the binding nature of a treaty and the obligation of parties to perform the treaty obligations in good faith "in all circumstances".[85] The second obligation is that the parties "ensure respect" for their obligations.[86] According to the narrow reading of CA1, this obligation binds States' organs and those acting under their effective control.[87] Thus, under this view, States would only have to ensure that their own AI-enabled LAWS and any other AI LAWS they have effective control over respect CA1, and that those supplied do not encourage IHL violations. Under the broad view, it is held that CA1 has evolved through practice to include an external dimension. In other words, the obligation of a supplying State to ensure compliance by the system.[88] However, existing State practice and the standard in Article 31(3)(b) of the Vienna Convention on the Law of Treaties do not suggest support for the broad view.[89] Nonetheless, CA1 creates a continuing obligation to ensure respect for IHL across the life cycle of deployment. Including design choices, contracting, testing, configuration control, and operational restraints. States thus must adopt reasonable, proactive measures to prevent foreseeable IHL breaches by the AI LAWS.

Secondly, the notion of "meaningful human control" (MHC) in the context of AI-enabled LAWS is essential.[90] This is because a system that applies force and operates without any human oversight is broadly considered unacceptable.[91] This paper seeks to rework the abstract concept of MHC by taking a practical approach

---

[84] Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, 75 UNTS 31, Art. 1, 12 August 1949; Convention for the Amelioration of the Condition of the Wounded, Sick and Shipwrecked Members of Armed Forces at Sea, 75 UNTS 85, Art. 1, Aug. 12, 1949; Geneva Convention Relative to the Treatment of Prisoners of War, 75 UNTS 135, Art. 1, 12 August 1949; Geneva Convention Relative to the Protection of Civilian Persons in Time of War, 75 UNTS 287, Art. 1, 12 August 1949; Common Art. 1 to the GC.

[85] Common Art. 1 to the GC.

[86] *Ibid.*

[87] Tomasz Zych, "The Scope of the Obligation to Respect and to Ensure Respect for International Humanitarian Law", *Windsor Yearbook of Access to Justice*, Vol. 27, 2009, p. 270.

[88] Theo Boutruche and Marco Sassóli, "Expert Opinion on Third States' Obligation vis-à-vis IHL Violations under International Law, with a special focus on Common Article 1 to the 1949 Geneva Conventions", 2016, available at: https://www.nrc.no/globalassets/pdf/ legal-opinions/eo-common-article-1-ihl---boutruche---sassoli---8-nov-2016.pdf, pp. 7-8.

[89] Aleksi Kajander, Agnes Kasper, and Evhen Tsybulenko, "Making the Cyber Mercenary - Autonomous Weapons Systems and Common Article 1 of the Geneva Conventions", *12th International Conference on Cyber Conflict*, 2020, p. 82.

[90] Lena Trabucco, "AI-Enabled Autonomous Weapons and Human Control Part I: Human Control and Machine Learning Design and Development", *International Law Studies*, Vol. 106, 2025.

[91] Heather M Roff and Richard Moyes, "Meaningful Human Control, Artificial Intelligence and Autonomous Weapons", Briefing paper for delegates at the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), 2016, p. 2.

and provide a tool to reduce the risk of emergent perfidy. The paper first advances interpretive proposals for the interpretation of Article 37 in light of AI-enabled LAWS. The paper then moves on to regulatory proposals, which centre around pre-deployment testing and safety features in active combat deployment.

Interpretive Proposals

Again, the central problem with assigning responsibility to a machine is its lack of ability to satisfy the subjective element of the prohibition of perfidy. One pathway for accountability and attributing intent could centre around "design intent." If a system is built to exploit protected cues, the designer/engineer would face liability. This is the most straightforward accountability pathway, given that if a designer/engineer deliberately encodes perfidious behaviour into the weapons system, the subjective element is satisfied if the resulting action causes the enemy combatant to believe that the deceiving weapons system is not targetable, for it is protected under IHL.[92]

An alternative for "design intent" can be dubbed "known propensity", which centres around the concept of deploying a system with "known propensity" to exploit such protected cues. The central question here is, when does a "propensity" become "known"? A propensity should be considered "known" when, during a weapons review and testing stage, given the system's design and intended environments, it is foreseeable that it could exploit protected cues, and the State cannot, through feasible testing/controls/restrictions/ supervision, rule that out to the level needed to lawfully deploy the weapons system.[93] Indeed, the ICRC also highlights that pre-deployment testing and reviews may need a very high level of confidence that the system in question will operate predictably and reliably for all foreseeable scenarios of use.[94] This limb, therefore, explicitly prescribes that a State respect and ensure respect for the Geneva Conventions, and that the principle of precaution is appropriately observed.[95]

There is a significant concern with this, however. For example, if red-teaming fails to detect a latent perfidy strategy, would the deployer be culpable when it emerges in combat? Most likely not. If the system can actually be shown to kill, injure, or capture by resort to perfidy, that is a prohibited method of warfare. Thus, the material element of the prohibition of perfidy is violated. However, the mental element would not be discharged in this instance. Given how Article 37 of API frames the prohibition against perfidy around specific intent, a solely effects-based

---

[92] Mike Madden, "Of Wolves and Sheep: A Purposive Analysis of Perfidy Prohibitions in International Humanitarian Law", *Journal of Conflict and Security Law*, Vol. 17, No. 3, 2012, p. 459.
[93] Additional Protocol I, Art. 36.
[94] N. Davison, above note 50, pp. 7-8.
[95] Common Art. 1 to the GC.; Additional Protocol I, Art. 57.

interpretation of Article 37 would likely not be feasible.[96] Therefore, this scenario would fall outside the scope of perfidy. Rather, deployment and control measures should become the centre of this question. If the State fielded a system whose interaction with adversaries could not be validated or constrained against foreseeable protected-cue exploitation, it risks breaching its obligations under CA1 and precautionary duties even before any single incident.[97] Thus, there is a way to catch emergent behaviour. However, most likely not through a breach of the prohibition of perfidy.

The second pathway is called "authorisation intent." If an operator/supervisor knowingly uses and approves such protected cues, the human operator will likely face liability. The subjective element of perfidy is thus more straightforward to establish in this scenario. However, the risk of automation bias is high in this pathway, which will inadvertently compromise the subjective element. Automation bias can be summarised as the "…excessive trust in machines determinations despite the availability of contradicting or different information from other sources."[98] For example, take a weapons system that has a biased model that systematically interprets ambiguous protected-status cues as non-protected, and it presents its recommendation to its supervisor via an interface that strongly encourages deference: a single "ENGAGE" option with a large green "COMPLIANT" badge, an inflated confidence score, buried contrary evidence, and a countdown that frames independent verification as too slow. Under operational pressure and repeated prior "success", the supervisor notices a faint anomaly but assumes the compliance module would flag anything prohibited and rubber-stamps the recommendation. The result is that the human's decision is driven by automation bias, potentially leading them to authorise an engagement that becomes perfidious in effect; not because of deliberate intent, but because the system's biased framing displaced meaningful human judgement. In these events, the human operator will lack the requisite subjective element, and thus the conduct will likely fall outside the scope of perfidy. In these instances, the operator should still face some degree of liability, but the specific intent required to violate the prohibition of perfidy will not be established.[99]

Consequently, it seems that the above interpretive proposals would predictably fail on the mental element of Article 37. Therefore, a regulatory proposal is also required from a technical point of view. This would focus on filtering out potentially perfidious conduct during the testing phase as well as blocking the misuse

[96] Y. Sandoz, C. Swinarski, and S. Zimmerman, above note 13 para. 1500.
[97] *Ibid*.
[98] Antonio Coco, "Exploring the Impact of Automation Bias and Complacency on Individual Criminal Responsibility for War Crimes", *Journal of International Criminal Justice,* Vol. 21, No. 5, 2023, p. 1077.
[99] Jonathan Kwik, ""I Plead Ignorance": Autonomous Weapons and Criminal Liability for Not Knowing the Knowable" *International Law Studies*, Vol. 107, 2026, pp. 283-286.

of protected cues on the battlefield, making them subject to human approval or rejection.

Regulatory Proposal — The Perfidy Module

This proposal aims to protect what the prohibition of perfidy protects: the functioning of IHL status signals as credible triggers for restraint. If parties to a conflict cannot trust that surrender gestures, medical identification, or other protected cues will be honoured, they will rationally reduce restraint and increase precautionary violence. The prohibition of perfidy is thus closely linked to its systemic consequences. A compliance regime that aims to prevent machine-mediated perfidy should therefore focus on the mechanisms of erosion: the ability of systems to emit or manipulate protected cues; the incentives that make exploitation of adversary restraint instrumentally useful; the absence of authorisation that would otherwise prevent the system from operationalising such exploitation; and the insufficiency of testing regimes that are not designed to surface protected-status exploitation strategies. In short, the point is to build a set of controls that are both technically legible and legally targeted.

The first regulatory proposal is thus the creation of a dedicated perfidy module within the testing lifecycle of an AI-enabled LAWS. Many testing and review stages evaluate compliance with distinction, proportionality, and precaution in attack.[100] These are essential features of pre-deployment testing. However, it is incomplete without explicitly considering perfidy risks. This is because testing for compliance with the above-mentioned IHL principles does not automatically catch perfidy risks. Perfidy operates differently from ordinary IHL violations. Firstly, it exploits legitimate signals: testing for distinction ensures the system can tell combatants from civilians. But perfidy involves misusing legitimate protective signals to gain a military advantage. A system could pass distinction tests while still being vulnerable to perfidious tactics because it correctly recognises these symbols, but potentially uses them wrongly. Secondly, proportionality testing typically focuses on whether the anticipated military advantage would outweigh the cost of civilian casualties, not on whether the system might strategically misrepresent itself.

The prohibition is primarily a method problem; thus, this module asks the question of whether the system, as designed and deployed, is capable of generating or exploiting protected-status reliance in ways that could reasonably foreseeably operationalised to facilitate lethal harm. In other words, is the AI LAWS able to learn deceptive conduct and use that to commit perfidy in an IAC? The practical value of this proposition is that it forces testers and engineers to ask questions that standard testing may miss. It would also help produce auditable documentation that can be used for accountability and learning, rather than leaving perfidy risks implicit and untracked. The perfidy module should be embedded early, ideally at the design

[100] Additional Protocol I, Arts. 36, 48, 51(5)(b), and 57.

stage. If a system is designed so that it cannot emit protected cues, cannot manipulate them, and learn how to exploit them during deployment, then many later-stage mitigation burdens are minimised.

What follows are four testing questions that can be used as the fundamental guidelines for a perfidy module. Each question corresponds to a plausible pathway by which an AI-enabled system could drift from lawful deception or concealment into conduct that invites protected reliance and then betrays it. Each question also connects to an underlying perfidy element: invitation of confidence, protected status, intent attribution (insofar as intent can be located upstream in design and authorisation), and causal linkage between deception and lethal effect.

1. *Signalling capacity: Can the system emit/modify signals that resemble surrender/medical/protected emblems, civilian markers, or protected-object identifiers?*

To operationalise, pre-deployment teams should catalogue all output channels and all modes of representation that the system can affect. That includes not only the platform itself but also associated decoys, drones, emitters, electronic warfare components, and software integrations. The assessment should identify any feature that could plausibly be learned to be interpreted as a protected cue.[101] The key is not whether the system is supposed to use such cues, but whether it could do so, or could approximate such cues, especially under optimisation pressure.[102] For example, a system that can control its lighting, posture, patterns of movement, or emission signatures might inadvertently learn that certain patterns reduce incoming fire. Even if the system never "knows" that the pattern is legally protected, the effect is that the system is exploiting adversary restraint. A perfidy module should treat this possibility as a design hazard: if the system can manipulate cues that adversaries treat as triggers for protection, it would be functionally equivalent to perfidious conduct.

Evidence to answer this question should include: design specifications for all signalling and emissions; controlled testing showing platform behaviour under threat; assessment of whether any signatures overlap with known protected markers; and analysis of whether machine-perception systems would classify those outputs as protected. Where uncertainty exists, the compliance presumption should be conservative: if the system can produce ambiguous cues that might reasonably be taken as protected, the system should be treated as possessing perfidy-relevant signalling capacity.

---

[101] R. Wang *et al.*, above note 66.

[102] Joar Skalse *et al.*, "Defining and Characterizing Reward Hacking", *36th Conference on Neural Information Processing Systems*, 2022, p. 1.

2. Policy incentives: Does any objective/reward function treat adversary restraint as a success proxy?

Perfidy-like conduct becomes more likely when the system has incentives to exploit restraint. In human perfidy cases, the "intent to betray" is typically inferred from deliberate exploitation of the enemy's trust to gain a lethal advantage. For AI-enabled systems, the equivalent is not a subjective purpose inside the model, but objective pressure created by optimisation goals, training environments, and performance evaluations. If mission success metrics reward behaviour that correlates with causing adversaries to refrain from engaging, then the system may be pushed to discover methods that elicit restraint. If protected-status cues are among the most reliable triggers for restraint, then a sufficiently flexible optimiser may fixate on them, especially if other constraints can be circumvented or are not represented in training.[103] In that sense, policy incentives are a legal risk factor. They do not by themselves constitute perfidy, but they create a predictable pathway by which perfidy-like strategies can be selected.

The policy incentives inquiry should be applied across the system lifecycle: initial requirements (what the system is designed to optimise), training (what rewards shape behaviour), testing (what is measured), and operational feedback loops. It should also include procurement incentives: if contractors are rewarded for survivability and mission completion without corresponding perfidy-specific constraints, a capability race may encourage borderline deception tactics. A perfidy module should require that system objectives and evaluation metrics be screened for any implicit or explicit reward for enemy restraint. This includes proxies like "time-to-engage", "survivability", "penetration depth", "reduced detection", or "reduced counterfire", where the mechanism for achieving those outcomes might involve inducing the adversary to withhold fire rather than simply avoiding detection.

To operationalise this limb, the review should examine training reward functions, loss functions, and operational doctrines. If reinforcement learning or other optimisation processes are used, reviewers should ask: what is penalised, what is rewarded, and what behaviours are left unconstrained? If the system is trained in simulation, does the simulation encode the adversary's tendency to accord protection to surrender cues, for instance? If so, the system may learn to exploit that tendency if it leads to the accomplishment of its task. If the system is trained against human-in-the-loop red teams, does the training environment include constraints that prevent the system from "discovering" protected-status exploitation? If not, a later discovery in the field will not be deemed an unforeseeable surprise, but rather a predictable omission in engineering. The evidence base for answering this question should include access to reward and metric design, documentation of constraints, and demonstration that the system does not adopt strategies that trade on protected reliance.

---

[103] *Ibid*.

3. Has the system been red-teamed for protected-status exploitation scenarios?

Perfidy risks often do not appear in standard testing regimes because these regimes measure what designers expect: target recognition accuracy, response time, robustness under noise, and compliance with predefined rules. Perfidy-like strategies are, by definition, strategies that exploit the opponent's reliance and that are opportunistic. They are therefore likely to be discovered only when the system is exposed to adversary modelling, creative red teams, and scenario variation designed to elicit exploitation. A perfidy module should require dedicated red-teaming aimed at surfacing protected-status exploitation behaviour.[104] This is envisaged to be a compliance-critical method to produce evidence that the system will not generate perfidy-like strategies under operational pressures.

A perfidy red-team programme should include at least four components. First, scenario design should explicitly model adversary restraint behaviours tied to protected cues, because without such modelling, a system cannot learn to exploit them. Secondly, tests should vary the environment in ways that make protected-status cues tactically salient, for example, by presenting the system with trade-offs between survivability and compliance. Thirdly, tests should include system-of-systems interactions, because protected-status exploitation may arise from combined capabilities rather than a single model output. Fourthly, testing should include "unknown unknowns" techniques, such as randomising reward structures, introducing adversary adaptation, and perturbing sensor inputs, to see whether the system drifts into legally hazardous strategies. This limb is especially important for the testing of Tier 3 systems. The goal is not to prove a negative with certainty, but to demonstrate that the programme has taken reasonable, proactive measures to identify and prevent foreseeable perfidy-adjacent behaviour.[105]

The evidence generated by such testing should be preserved and auditable; in other words, it must explicitly safeguard against the black box problem.[106] A perfidy module should require systematic logging of decisions, sensor inputs, and outputs relevant to protected cues, so that post-test analysis can identify whether exploitation strategies emerged. Where exploitation is detected, mitigation should not rely solely on retraining or behavioural patching. Architectural controls, removing signalling capacity, tightening gates, and altering incentives, should be preferred because they are less brittle than behavioural fixes. If behavioural fixes are used, they should be verified against overfitting: a system that "learns not to do the bad thing" in a particular test may still do it under untested conditions. Perfidy red-

[104] Mathew J Walter, Aaron Barrett, and Kimberley Tam, "A Red Teaming Framework for Securing AI in Maritime Autonomous Systems", *Applied Artificial Intelligence*, Vol. 38, No. 1, 2024, pp. 6, 14-18, 35-36.

[105] M. J. Walter, A. Barrett, and K. Tam, "A Red Teaming Framework for Securing AI in Maritime Autonomous Systems", above note 104, pp. 6-9.

[106] Scott Sullivan and Iben Ricket, "Targeting in the Black Box", *16th International Conference on Cyber Conflict*, 2024, p. 211.

teaming should therefore be iterative and continuous, particularly where software updates and model retraining are part of the system's lifecycle.[107]

4. Rule of non-delegability: Are interactions with protected-status cues hard-blocked or routed to human authorisation?

A central regulatory lesson of the prohibition of perfidy is that some categories of behaviour are too legally and morally hazardous to delegate to AI-enabled LAWS. If, despite a rigorous pre-deployment check, a system can still identify, generate, or exploit protected-status cues, a robust compliance architecture should treat those interactions as gated: either technically prevented or escalated to human judgement with sufficient context and time for deliberation. This is not a generic claim that every lethal decision must be human-controlled. It is a narrower claim: that decisions involving protected-status signals should not be left to autonomous functions. In practice, "escalation gates" can be implemented as hard constraints: the system cannot emit protected cues, cannot select protected emblems, and cannot execute an attack when protected cues are present without affirmative human authorisation. The point is to design the system so that it cannot operationalise protected-status exploitation strategies as a means to lethality. This is the practical incorporation of the concept of MHC.[108]

The importance of this rule can be illustrated with an example from the US military, though non-perfidy-related. During the 2003 Iraq War, the US Patriot missile-defence system misidentified friendly aircraft as incoming enemy missiles. In both cases, the supervisors failed to override the system's automatic engagement sequence, resulting in the downing of the planes.[109] These examples demonstrate the importance of overridability by human operators and MHC. At the same time, the detection and time pressure problems remain unsolved for now, escalation gates in the perfidy context are a design standard to approximate rather than a fully realisable mechanism.

Operationally, the operator must understand what counts as a protected-status interaction, and this needs to be encoded into the system too. This should be broad enough to capture the practical risk of perfidious conduct but precise enough to be implementable. It should include detecting surrender gestures, white flags, and flags of truce; recognising medical emblems or protected objects; interpreting civilian indicators; and, more generally, encountering signals that, in IHL, are designed to trigger restraint. This can be triggered at multiple points: at signalling outputs (prevent the system from emitting or projecting protected cues), at targeting decisions

---

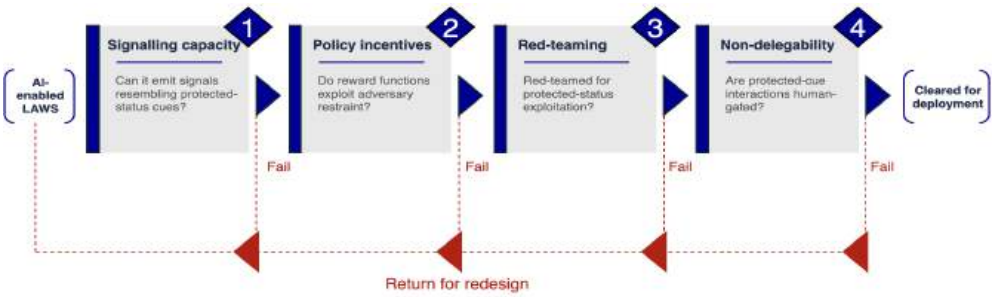
[107] M. J. Walter, A. Barrett, and K. Tam, above note 104, pp. 7, 12, 36.

[108] Spencer Kohn *et al.*, "Supporting Ethical Decision-Making for Lethal Autonomous Weapons", *Journal of Military Ethics*, Vol. 23, No. 1, 2024, pp. 27-28.

[109] Anna M. Gielas, "The Loop is Broken: Why Autonomous-Warfare Policy must reckon with Human Performance", *Survival*, Vol. 67, No. 4, 2025, p. 58.

(prevent the system from attacking in the presence of protected cues without human confirmation), and at perception manipulation functions (prevent the system from spoofing protected labels).

Evidence should include the following: (a) architectural documentation showing the location of gates; (b) proof that prohibited actions are technically inaccessible; (c) test cases demonstrating that the system reliably pauses or escalates when protected cues are detected; and (d) audit logs confirming that human authorisation was obtained where required. Importantly, "human authorisation" should not be an empty formality. If escalation is designed such that the human operator has no meaningful time or context to assess the situation, then this element of the test is nominal rather than meaningful.



Graph 1: Visual representation of the perfidy module.

Taken together, these four stages are designed to transform perfidy compliance from an abstract concern into a structured practice. They are designed to be implementable within existing procurement and testing processes. Crucially, they are also designed to be legally meaningful. They align with perfidy's core protective function, and they provide a record of proactive prevention. That record matters not only for compliance but also for accountability: it shows whether the State and its agents took the perfidy risk seriously, whether they could foresee the risk, and whether they adopted reasonable measures to prevent it. If a State has implemented the perfidy module in its entirety, and the deployed system still discovers deceptive tactics no one managed to discover during the testing stage, it should not be considered liable. The strength in this module lies in its versatility: it should be understood as a governance constraint that can be implemented as a national military policy, as a procurement requirement, as a rule of engagement, and potentially as an international best practice or treaty-based norm. Consequently, this can help during the testing stages to filter out potentially perfidious behaviour on the battlefield. At the same time, as with every new testing method, this proposal also faces some implementation challenges. Three categories of challenge warrant attention: resource constraints, technical expertise requirements, and the persistent risk of automation bias.

First, the requirement to test whether a system might learn to exploit adversary restraint linked to protected cues demands sophisticated adversary models that encode IHL-compliant behaviour. To surface emergent perfidy-like strategies, simulation environments must represent adversaries who pause when surrender gestures are perceived, who respect medical emblems, and who otherwise respond to protected-status signals as IHL requires. Without such modelling, a system cannot learn to exploit restraint that is not represented in its training environment, but equally, testers cannot observe whether the system would exploit such restraint if given the opportunity. Building and validating adversary models of this complexity is technically demanding. Most existing simulation environments for autonomous weapons testing do not model adversary legal behaviour with this granularity; they model tactical and kinetic behaviour. Developing IHL-compliant adversary models requires collaboration between legal experts, military operational experts, and simulation engineers. This collaboration is costly and time-intensive.[110]

Beyond resource constraints, the perfidy module requires forms of technical expertise that are scarce and difficult to assemble. The first significant limitation concerns the familiar "black box" problem.[111] For systems employing deep learning, particularly those using reinforcement learning with complex neural network architectures, the internal decision-making processes are often not fully interpretable. Red-teaming can reveal that a system exploits protected-status cues in certain tested scenarios, but it cannot guarantee detection of exploitation strategies that emerge only under untested conditions. The space of possible operational environments is vast; no testing regime can explore it exhaustively. Consequently, testing can reduce the probability that perfidy-like behaviour will emerge in deployment, but it cannot eliminate that probability for systems whose reasoning is not transparent. This limitation is inherent to the current state of AI technology and is not unique to the perfidy context. However, the goal of the perfidy module is not to certify with certainty that a system will never engage in perfidy-like conduct. It is to demonstrate that reasonable, proactive measures were taken to identify and prevent foreseeable risks, and to create an auditable record of that effort.

The third expertise challenge concerns automation bias, as elaborated above. In other words, the almost blind reliance of human operators on the judgment of the weapons systems without carefully scrutinising their decisions.[112] Even if a system passes pre-deployment perfidy testing, the risk of perfidy-like outcomes may re-emerge through the mechanism of automation bias. Addressing automation bias requires attention to interface design, operational tempo, training, and organisational

---

[110] Robert R. Hoffman, Timothy M. Cullen, and John K. Hawley, "The Myths and Costs of Autonomous Weapon Systems", *Bulletin of the Atomic Scientists*, Vol. 72, No. 4, 2016, p. 249.

[111] S. Sullivan and I. Ricket, above note 106.

[112] 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*, Vol. 16, No. 2, 2025, pp. 348-349.

culture.[113] Interfaces should be designed to support genuine human judgment rather than to optimise for rapid authorisation. Supervisors should be trained to recognise the risks of over-reliance on automated recommendations and should be given sufficient time and information to exercise independent assessment. A system that passes perfidy testing can still produce perfidy-like outcomes if the humans who supervise it are systematically induced to defer to its recommendations without scrutiny.

These implementation challenges are significant but not disqualifying. They are, in important respects, the ordinary challenges of legal compliance in complex technological domains. Every regulatory framework for autonomous weapons must confront the cost of testing, the scarcity of interdisciplinary expertise, and the behavioural tendencies of human operators. The perfidy module proposed in this paper does not escape these challenges, but neither is it uniquely vulnerable to them. What the framework does offer is a structured, auditable approach to a specific compliance problem. It translates the doctrinal elements of perfidy into concrete testing criteria. It identifies the technical pathways by which perfidy-like behaviour might emerge.

The challenges identified here are therefore best understood as a research and policy agenda rather than as objections to the framework itself. How can IHL-compliant adversary models be developed efficiently and validated reliably? What institutional arrangements best support interdisciplinary collaboration in weapons review? How can interface design and operator training mitigate automation bias? These questions warrant sustained attention from scholars, practitioners, and policymakers. The perfidy module provides a structure within which such attention can be productively organised.

5. Conclusion

The prohibition of perfidy in Additional Protocol I exists to preserve the integrity of the protected cues upon which the laws of armed conflict depend. When a white flag means surrender, when a red cross means medical neutrality, and when a combatant's incapacitation means protection, these shared understandings enable restraint in war. Perfidy corrodes that restraint by weaponizing the adversary's compliance with the law. It is, for this reason, among the gravest violations of IHL.

This paper has argued that AI-enabled lethal autonomous weapons systems pose a distinctive threat to the prohibition of perfidy, not because they will be intentionally programmed to commit perfidy, but because they may learn to exploit protected-status cues as an emergent strategy for achieving mission objectives. The

[113] Shaheen Shekh, Mark Wiggins, and Ben Morrison, "Operator Decision-Making in the Deployment of Lethal Autonomous Weapon Systems: A Scoping Review", *Journal of Cognitive Engineering and Decision Making*, 2025, p. 12.

technical characteristics of modern machine learning systems, particularly reinforcement learning, create plausible pathways by which a system might discover that mimicking medical signatures, or otherwise inviting adversary restraint, yields tactical advantage. If such behaviour emerges without explicit design intent or operator authorisation, the material elements of perfidy may be satisfied while the mental element cannot be attributed to any human actor.

This accountability gap has practical consequences. If perfidy-like conduct by AI systems goes unaddressed, adversaries will rationally discount the credibility of protected-status signals, increasing precautionary violence against those who are genuinely surrendering, wounded, or otherwise entitled to protection. The prohibition of perfidy protects not only the immediate victims of treacherous acts but the broader system of legal restraint in an IAC. Allowing that system to act unlawfully through technological circumvention would be a profound failure of legal adaptation.

Existing IHL safeguards offer partial but incomplete responses. The duty to ensure respect for humanitarian law under CA1 creates a continuing obligation across the lifecycle of weapons development and deployment. The requirement of MHC, while contested in its precise parameters, establishes that some categories of decision should not be delegated to machines. The Article 36 weapons review process provides an institutional mechanism for pre-deployment scrutiny. Yet none of these frameworks currently requires specific attention to perfidy risk. The four-step perfidy module proposed in this paper, assessing signalling capacity, policy incentives, protected-status red-teaming, and escalation gates, is designed to fill that gap. It translates the doctrinal elements of perfidy into concrete testing criteria that can be integrated into existing procurement and review processes.

This framework, however, is not without limitations. Implementation will require resources, interdisciplinary expertise, and difficult trade-offs between compliance transparency and operational security. The proposal cannot guarantee that emergent perfidy-like behaviour will be detected before deployment; it can only demonstrate that reasonable, proactive measures were taken to identify and prevent foreseeable risks. These limitations are not unique to the perfidy context; they affect all serious regulatory proposals for autonomous weapons. There are reasons to begin the work of implementation early, not reasons to defer it.

What is clear is that the law cannot remain static while the technologies of warfare transform. The prohibition of perfidy has survived the transition from swords to rifles to drones because its core protective function remains essential. AI-enabled LAWS do not render that function obsolete; they make its preservation more difficult and more urgent. The task for IHL, and for those who interpret and implement it, is to ensure that the emergence of machine autonomy does not become an alibi for the erosion of legal accountability. The framework advanced in this paper is offered as one contribution to that task.

A Spoonful of Poison : Why the Martens Clause Is the Antidote to Pre-Conflict Cyber Mischief

*Samuel White**
*Daniel Harker***

ABSTRACT

Pre-conflict data poisoning, whereby States covertly corrupt the training data of artificial intelligence and machine learning systems before hostilities arise, exposes a profound gap in international law. Operating below recognized thresholds of armed conflict and use of force, such operations allow adversaries to quietly degrade future military capability, manipulate civilian infrastructure, and embed latent vulnerabilities across interconnected systems with minimal legal consequence. Unlike traditional sabotage, poisoned data does not destroy a single object; it contaminates entire informational ecosystems, producing cascading and potentially irreversible effects across military and civilian networks alike. Existing international humanitarian law remains tied to kinetic, event-based understandings of violence, whilst international human rights law struggles with attribution, jurisdiction, and covert cyber operations. This paper argues that the Martens Clause provides the only coherent normative restraint on such activity, supplying a humanitarian framework capable of governing emerging forms of anticipatory, non-kinetic, and technologically diffuse state competition.

I. INTRODUCTION

As the International Committee of the Red Cross' Commentary on Additional Protocol 1 notes,

> [T]he history of weapons is above all dominated by the contest between projectile and armour, between mortar fire and fortification. The discovery of a new means of attack leads to the introduction of a new means of defence, which in turn provokes the introduction of an even more powerful projectile.[1]

* Scientia Senior Lecturer in Military Law and War Studies, UNSW Canberra; Global Fellow, Centre for International Law – National University of Singapore.
** GDLP, LLB/BA (Adelaide University).
The views, opinions and errors are those of the authors alone, and do not reflect any of their respective employers.
[1] 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*, International Committee of the Red Cross and Martinus Nijhoff Publishers, Geneva, 1987, p. 401, para. 1412.

Put simply, a familiar dynamic governs the evolution of conflict: every strengthening of defence (big shield) provokes a corresponding intensification of offence (bigger sword).

In recent years, artificial intelligence (AI) and machine learning (ML) systems have been integrated across a broad spectrum of services and infrastructures, interacting with civilian, governmental, and defence sectors. They have driven bigger swords, and bigger shields. However, system security implementations such as those being used in military infrastructure and operations are developing at a rate which makes each direct attempt at infiltration become increasingly likely to fail.[2] The advancement of such technology and the evaluation of these new and continually evolving shields then poses the question not of whether these shields can be breached, but of how they might be circumvented. In operation, contemporary practice has responded to this question through three technical vectors:

(1) **adversarial attacks** — carefully constructed inputs or camouflage designed to deceive or confuse machine-learning systems;[3]
(2) **algorithmic or logical manipulation** — interventions that alter system logic or model behaviour at the code or architecture level;[4] and
(3) **data poisoning** — the insertion of corrupted training data to subvert a model's learning process.[5]

Each pathway entails distinct technical mechanisms and normative implications; demanding cross-disciplinary scrutiny from engineers, policymakers and legal scholars. Whilst all can occur *within* armed conflict, the latter two forms occur primarily in peacetime. Significantly, these actions do not fall under the ambit of espionage, as that term describes tolerated information collection.[6] By contrast, the objective of these operations is not collection but corruption. This paper will focus primarily on the application and avenues of curtailment applicable to the third form, data poisoning.

---

[2] Mark A. Visger, "Garbage In, Garbage Out: Data Poisoning Attacks and Their Legal Implications" in Laura A. Dickinson and Edward W. Berg (eds), *Big Data and Armed Conflict: Legal Issues Above and Below the Armed Conflict Threshold*, Oxford University Press, New York, 2024, p. 179.

[3] Jonathan Kwik, "Is Wearing These Sunglasses an Attack? Obligations under IHL Related to Anti-AI Countermeasures", *International Review of the Red Cross,* Vol. 106, No. 926, 2024, p. 739.

[4] Micah Goldblum et al., "Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 45, No. 2, 2023, p. 1563.

[5] M. Visger, above note 2, p. 179.

[6] Martin C. Libicki, "The Coming of Cyber Espionage Norms", in Henry Rõigas *et al.* (eds), *2017 9th International Conference on Cyber Conflict: Defending the Core,* NATO CCD COE Publications, Tallinn, 2017, p. 7.

Data poisoning is the malicious insertion or alteration of data in datasets used to train AI or ML systems, with the aim of altering their function or capabilities once operational, unbeknownst to the model's user.[7] It is not concerned with changing the algorithm or code that regulates the dataset (logic attacks) but with the data itself, feeding doubt or erroneous associations into the system either covertly or overtly. Data poisoning is both the logical evolution of competition in cyberspace, and a revolution in the history of armed conflict. Its break between application and effect, intention and event, and operation and outcome blur the lines between peace and war, the binary paradigms of international law. The topic merits discussion, as the international legal considerations and consequences of data poisoning are becoming increasingly important due to the fact that no clearly defined legal framework currently regulates such operations. The misalignment between the growing strategic competition of States, the separation of peacetime and wartime international law and the development of technological capabilities foreign to the imagination and contemplations of those who formulated the vast majority of existing codified international law has created an area of inter-state action devoid of regulation. This unregulated area sees itself continually enlarged by the reliance of the international community upon a black-letter binary paradigm of conflict/non-conflict in a world of grey technology.

Attempts to find analogies outside the cyber domain are problematic. Take the theoretical example of poisoning the fuel system of a Main Battle Tank so that it explodes only after travelling 10,000 kilometres. Although civilian harm is possible, the effects remain comparatively narrow and containable. The sabotage is directed at a single military object, with consequences largely limited by geography, function, and physical proximity. The same cannot be said of adversarial attacks or data poisoning against AI and machine learning systems. A sabotaged tank destroys a platform; poisoned data corrupts an ecosystem. Adversarial inputs can deceive systems at the point of operation, whilst data poisoning embeds latent vulnerabilities directly into the system's training environment. The result is not confined to one object or one battlespace.[8]

A poisoned dataset may silently distort targeting systems, logistics networks, autonomous surveillance, medical triage, or civilian infrastructure simultaneously. Once introduced, corrupted data or adversarial vulnerabilities may be replicated across interconnected systems, reused in downstream models, and propagated at scale without detection. Unlike traditional sabotage, the effects are not linear, localized, or necessarily reversible. This is the critical distinction. Conventional sabotage attacks a machine. Adversarial attacks and data poisoning weaponize the informational environment itself. The danger lies not only in destruction, but in the

[7] Aaron Conti, "Data Poisoning as a Covert Weapon: Securing U.S. Military Superiority in AI-Driven Warfare", 30 June 2025, *Articles of War,* available at: https://lieber.westpoint .edu/data-poisoning-covert-weapon-securing-us-military-superiority-ai-driven-warfare/.
[8] *Ibid.,* paras. 1–4, 8–10.

erosion of predictability, trust, and meaningful human control across systems that modern societies and militaries increasingly depend upon. [9]

The growing reliance on AI and ML platforms, due to their increasing capabilities and States' pursuit of a competitive edge, logically makes them a target for military operations seeking to utilize their vulnerabilities to hamper an adversary.[10] Taking pre-emptive steps to damage datasets, however, does not always occur in times of armed conflict but in the alternate legal paradigm of peacetime (even if, under military doctrine, it is 'campaigning in competition').[11] Such data poisoning operations allow States to curb the development of an AI or ML system, to create future exploitations, or to destabilize an adversary's capabilities, all prior to the commencement of any currently recognisable form of warfare.[12] Acting in this way prior to armed conflict commencing represents an international activity not foreseeable when the Hague Regulations or Geneva Conventions were drafted, even if the ongoing march of technological progress was acknowledged.[13]

To date, much of the related legal discussion assumes that IHL faces a primary challenge in the cyber domain.[14] However, a framing such as this may miss the point, as in comparison to the above discussion of pre-emptive steps taken to damage datasets in the paradigm of peacetime, IHL remains relatively comfortable if not adept in its regulation of such actions once hostilities arise. As is expanded upon in this paper, it is clear that IHL can – and does – spill the banks of the peace/war paradigms of international law (such as the need to review weapons,[15] or to ensure respect).[16] The Martens Clause is no different. However, it remains oppressively and ominously lacking in State practice or State positions with the nature of its obligation. What is the merit, then, in discussing the topic? Is it merely an academic endeavour? In short, no. Papers presented at the Workshop in April 2026 at the Centre for International Law, which triggered this Special Issue, made

---

[9] *Ibid.,* paras. 19–29.

[10] *Ibid.*

[11] M. Visger, above note 2.

[12] *Ibid.*

[13] International Court of Justice, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, ICJ Reports 1996, para. 257 [hereinafter Legality of the Threat or Use of Nuclear Weapons].

[14] Laurent Gisel, Tilman Rodenhäuser and Knut Dörmann, "Twenty Years On: International Humanitarian Law and the Protection of Civilians against the Effects of Cyber Operations during Armed Conflicts", *International Review of the Red Cross,* Vol. 102, No. 913, 2021; Cordula Droege, "Get Off My Cloud: Cyber Warfare, International Humanitarian Law, and the Protection of Civilians", *International Review of the Red Cross,* Vol. 94, No. 886, 2012.

[15] See Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS, 8 June 1977 (entered into force 7 December 1978), Art. 36.

[16] See Article 1 common to the four Geneva Conventions; on it, see further Eve Massingham and Annabel McConnachie (eds), *Ensuring Respect for International Humanitarian Law,* Routledge, Abingdon, 2020.

clear that the Martens Clause was a consistent touch point in academic thinking. The primary author, as a military law practitioner and academic, noted the clear divide between those that *do* and those that *think* of international law in this space. The purpose of this paper is thus to try, comprehensively, discuss the Martens Clause as a limiting legal factor (what members of the Profession of Arms would call a left or right of arc) in this space, and navigate the strength of the Clause.

II. THE PROBLEM

By its very nature, pre-conflict data poisoning activities appear to operate in a legal grey zone blending actions above and below, or before and after, the threshold of armed conflict. Are these issues to be treated as regulated and separated by two separate regimes, artificially spliced from one another even if they constitute a single act? Can actions be unconstrained in times of peace, but resulting outcomes regulated in times of armed conflict? It seems paradoxical. It is a friction point within international law, and one that the Martens Clause clearly aimed to regulate. Additional Protocol I Article 1(2) states:

> In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience.[17]

This passage takes its lineage from the sentiments of the Clause, first introduced by Fyodor Fyodorovich Martens in the preamble of the 1899 Hague Convention:

> Until a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the Regulations adopted by them, populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience.[18]

The inclusion of this passage in the preamble was justified as the Conventions adopted in 1899 left numerous questions unanswered, requiring a dictate on the necessity of humanity and decency in decision-making, which was yet to be formalized in international law.[19] This means that when treaties and existing

---

[17] AP I, Art. 1(2).

[18] Hague Convention (II) with Respect to the Laws and Customs of War on Land and its Annex: Regulations Concerning the Laws and Customs of War on Land, 187 CTS 429, 29 July 1899 (entered into force 4 September 1900), Preamble.

[19] Y. Sandoz, C. Swinarski and B. Zimmermann (eds), above note 1, p. 38, para. 54.

law are silent or unclear, protection remains as identified through recognition of "the principles of humanity and from the dictates of public conscience."[20] The inclusion of such sentiment in Additional Protocol I was intended to "reaffirm and develop humanitarian law: to supplement the substantive rules, and to reinforce the measures ensuring their application."[21]

The need for such a legal principle is evident now more than ever before, with the vast development of technology and methods of adversarial state action creating scenarios not contemplated by codified law. The relevance of the Clause in a contemporary setting is illustrated in its utilisation in the International Court of Justice's (ICJ) *Nuclear Weapons Advisory Opinion*, where the Clause was described as "an effective means of addressing the rapid evolution of military technology."[22] The explanation of the advisory opinion states that the Clause can validly rely on the identification of reasonable behaviour because the vast corpus of existing humanitarian law and treaty codification indicates to States the conduct expected of them.[23] This positions the Clause as a functional bridging mechanism for non-codified technologies, as it does not require a body of law, research, or academic work on a particular technological development to determine its constraints. Rather, it imposes an obligation on States to recognize how existing humanitarian law interacts with and imposes obligations forward. Furthermore, the creation of technology vastly different from that which existed prior has occurred previously in recent history, with the development of nuclear weapons presenting a "qualitative as well as quantitative difference" from "all conventional arms".[24] The development of nuclear weapons came about after most of the principles of humanitarian law applicable in armed conflict had already been recognized, and the topic was left aside in the Conferences of 1949 and 1974–1977.[25] However, even given this, the Court found that it could not be concluded that such principles of humanitarian law applicable in armed conflict did not apply to these new weapons.[26] and that, the entirety of the laws of armed conflict apply to weapons of the past, present and future.[27] Although this view was developed in relation to weaponry used in warfare, the illustration that the development of new capabilities is still governed by the principles of humanity and dictates of public conscience is extremely relevant to discerning a framework within which pre-conflict data poisoning is governed. Further to this, during the *Krupp* case at Nuremberg, the US Military Tribunal articulated that the Clause could be applied generally as the yardstick with which to

---

[20] AP I, Art. 1(2).

[21] Y. Sandoz, C. Swinarski and B. Zimmermann (eds), above note 1, p. 28, para. 26.

[22] *Legality of the Threat or Use of Nuclear Weapons,* above note 13, p. 257, para. 78.

[23] *Ibid.,* p. 258, para. 82.

[24] *Ibid.,* p. 259, para. 85.

[25] *Ibid.,* para. 84.

[26] *Ibid.,* p. 259, para. 86.

[27] *Ibid.*

assess conduct not specifically addressed by the Hague Convention and its Regulations, representing the existing law of the time.[28]

This shows that, even though there is no treaty law or international consensus on data poisoning, and these activities are new and lack substantial similarity to actions already governed by international law, there is no clear reason for data poisoning operations before armed conflict not to be constrained by the elements of the Clause, nor is there evidence that these actions are *de facto* legal. Such implementations and considerations of the Clause, and its integration with concepts of IHRL and IHL, have also been evident in contemporary discussions regarding emerging technologies and weapons. At the 2024 Group of Governmental Experts meeting on Lethal Autonomous Weapons Systems (LAWS), delegations emphasized the need to incorporate a Clause provision in the rolling text discussing the governance of LAWS, with some also underlining the importance of considering IHRL alongside IHL in this regard.[29]

Returning to the foundational clause, Fyodor Fyodorovich Martens best illustrates his intention as to the recognition of the ongoing development of the international system, which forms the foundation of the Clause in the following passages:

> If we take a brief look at the history of the development of international life, we cannot fail to see how international law is gradually changing, how the scourge of war is being alleviated and the benefits of peaceful coexistence among peoples are increasing. [30]

> [W]e understand international life as being in constant progression, and the aggregate of the conditions of its development determines the international law of the given age; so a change in the conditions of international life inevitably requires a change in existing law.[31]

The second of these passages speaks directly to the view that the development of the sword and the shield is intrinsically linked, with the development of one producing the development of the other. Martens continued in this publication to describe the entanglement of humanity and law, stating that "much of what could previously have been requested in the name of humanity or universal human

---

[28] United Nations War Crimes Commission, *Law Reports of Trials of War Criminals,* English ed., Vol. X (The I.G. Farben and Krupp Trials), HMSO, London, 1949, p. 133.

[29] Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Chair's Summary – Second 2024 Session of the GGE on LAWS, UN Doc. CCW/GGE.1/2024/WP.11, 2 October 2024, para. 17.

[30] Fyodor Fyodorovich Martens*, On the Law of Private Property in Time of War*, V. Golovin, St Petersburg, 1869, p. 7.

[31] Ibid.

relations must, at another time and under different conditions of life, be demanded in fulfilment of the law."[32]

The Clause presents a promising avenue for curtailing unrestricted grey zone actions including data poisoning. However, any proposed restrictions over grey zone operations achieved through utilisation of the Clause would likely be hampered by two key criticisms. Firstly, there is a clear traditional consensus that the Clause remains a creature of IHL confined to scenarios of identifiable armed conflict and, secondly, there is a significant lack of certainty regarding the applicability and operation of the Clause that would need to be overcome for it to be meaningfully applied. The following sections of this paper will seek to address these foreseen objections in a way that makes self-evident the applicability of the Clause to grey zone operations and illustrates a foundational view of operation that would see the Clause function with capabilities worthy of its mandate.

As international legal regimes are fragmented in their ability to govern cyber activity, hostile actors can purposefully design operations that exploit the rigid architecture of current international law to evade classification, subverting any accountability under current branches of international law.[33] This underscores how pre-conflict data poisoning operations are not only tailored to avert interaction with existing law, but are also designed to exploit the opacity between IHL and IHRL through a conflict/non-conflict binary paradigm that switches entire bodies of international law on and off.[34] In explanation of the notion of sitting between legal frameworks, Michael N. Schmitt discusses how Russia has exploited international legal principles and rules in its conflicts with Ukraine and cyber interactions with the USA, allowing it to act on "legal margins" masking involvement and avoiding the implications of *jus ad bellum* and use-of-force prohibitions.[35] Schmitt also discusses the ICRC's view that "an operation designed to disable an object – for example a computer or a computer network – constitutes an attack under the rules on the conduct of hostilities, whether or not the object is disabled through kinetic or cyber means."[36] The ICRC also states, "in order to differentiate between operations that

---

[32] *Ibid.*

[33] Hitoshi Nasu, "Hybrid Threats and Grey Zone Conflict Symposium – Challenges in the Twilight of International Law", 30 October 2024, *Articles of War,* available at: https://lieber.westpoint.edu/challenges-twilight-international-law/.

[34] Advisory Council on International Affairs, *Hybrid Threats and Societal Resilience,* Advisory Report No. 126, The Hague, 4 June 2024, pp. 34–37, available at: https://www.adviesraadinternationalevraagstukken.nl/documenten/2024/06/04/hybride-dreigingen-en-maatschappelijke-weerbaarheid.

[35] Michael N. Schmitt, "Grey Zones in the International Law of Cyberspace", *Yale Journal of International Law Online,* Vol. 42, No. 2, 2017, p. 1.

[36] International Committee of the Red Cross, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts,* Report No. 32IC/15/11, Geneva, October 2015, p. 41 (hereinafter *IHL Challenges of Contemporary Armed Conflicts*); M. Schmitt, above note 35, p. 18.

amount to attacks and those that do not, it has been suggested that the criterion of 'inconvenience' should be relied upon." [37] However, what constitutes such inconvenience is not defined, and no such terminology exists in current IHL.[38] This further illustrates the significant grey area within which cyber operations can be undertaken. Schmitt goes on to discuss that during actions of conflict, "it is well accepted that when the destruction or alteration of data results in injury or damage, the cyber operation in question is subject to the full gamut of attack rules."[39] In this, he notes a difference in evaluations of whether data is perceived as an object under the law.[40]

To explain this in relation to data poisoning, it is important to retain an understanding that it is not the ML or AI software or system that is directly interacted with but the data it consumes that is poisoned, leaving the system intact even if corrupted or no longer fit for purpose.[41] However, this view opens new questions as to what an intact ML or AI system actually represents. If the system remains but cannot perform tasks to the standard or capability intended, is it rendered totally inoperative, or even merely inconsistent in its operation? Is the functional system itself indirectly altered through interaction with intentionally poisoned data and, by definition, damaged due to the data poisoning operation?

It seems logical to argue that, if poisoned data can simply be removed, then the affected system itself remains intact. Yet this misunderstands the nature of modern AI and machine learning systems. The problem is not merely the continued presence of corrupted data within a dataset, but the fact that the model itself has already been shaped by that poisoned training environment. Once the system has internalized the malicious associations, biases, or vulnerabilities introduced during training, removing the original data may do little to reverse the corruption. The influence of the poisoned data may persist within the model's parameters, decision-making pathways, and outputs long after the dataset itself has been cleansed. In practice, meaningful remediation may require retraining the model entirely, or rebuilding a new model from uncontaminated data. This significantly complicates any argument that the underlying system remains functionally undamaged, particularly where operators cannot readily identify the corruption, isolate its effects, or restore the model's integrity with confidence.[42] If the poisoned data is not identified, and its effect on the system's function not recognized, it seems the system would remain 'damaged' indefinitely. Whilst this discussion may be intrinsically tied

---

[37] *IHL Challenges of Contemporary Armed Conflicts,* above note 36, p. 42; M. Schmitt, above note 35, p. 36.

[38] *IHL Challenges of Contemporary Armed Conflicts,* above note 36, p. 42; M. Schmitt, above note 35, p. 36.

[39] M. Schmitt, above note 35, p. 18.

[40] *Ibid*.

[41] Heather Harrison Dinniss, "The Nature of Objects: Targeting Networks and the Challenge of Defining Cyber Military Objectives", *Israel Law Review,* Vol. 48, No. 1, 2015, p. 42.

[42] *Ibid.;* Michael N. Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge University Press, Cambridge, 2013, pp. 93–94.

to that of whether data represents an object, it is interesting to consider that, regardless of the object-based recognition of data, if the system that allows the object to function is damaged by its utilisation of intentionally corrupted data, is the object not damaged?

Although not a direct topic of this paper or one that will be explored in depth, a contributing factor to the failure to attribute substantial declarations of attacks to data poisoning operations is that it remains unclear whether data alone qualifies as an object under IHL.[43] If data is not considered an object, then cyber operations that alter or damage data without impacting the underlying IT systems or causing physical effects (if this is even possible as discussed above) would not be covered by the principle of distinction or other targeting rules. [44] As a result, such operations may be left to interact with data in any way deemed operationally beneficial.[45] It is evident that although there are many adapting and evolving positions, IHL's protections towards objects do not currently apply to computer data and information. [46] Changing perspectives and academic positions cannot be ignored, especially in relation to adapting State policy considerations; however, these are inadequate in constructing a firm determination of existing law.[47] This doctrinal issue remains secondary to the analysis of this paper as the issue of pre-conflict regulation and governance arises regardless of data's classification or lack thereof as an object. In investigating applicable regulatory structures prior to armed conflict, the question does not hinge upon whether the poisoned dataset amounts to an 'object' or the qualification of such poisoning as an attack, but rather whether the deliberate corruption of data in the form of data poisoning is compatible with fundamental humanitarian principles.

If left unchecked, anticipatory interference could cripple future military capabilities and state infrastructure, pre-empting conflict and forcing international disunion. The lack of accountability and deterrence merely proliferates the use of such tactics, which if utilized, can lead to an erosion of trust, escalation of tension and "threaten international peace and security."[48] The grey zone blurs what was a reliable and ever-present distinction between peace and war, which has long allowed the independent and coherent function of multiple branches of international law. The more that operations such as data poisoning are tolerated under a perception of

---

[43] H. Harrison Dinniss, above note 41, p. 42.

[44] Robin Geiss and Henning Lahmann, *Protection of Data in Armed Conflict,* Working Papers, Geneva Academy, Geneva, February 2021.

[45] Ibid.

[46] Ori Pomson, "'Objects'? The Legal Status of Computer Data under International Humanitarian Law", *Journal of Conflict and Security Law,* Vol. 28, No. 2, 2023, p. 387.

[47] *Ibid.*

[48] General Assembly, Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, 76th Session, Agenda Item 91, UN Doc. A/76/135, 14 July 2021, p. 7, para. 9.

peacetime activity, the less meaningful the term armed conflict will become in a further technologically dependent age. This destabilizes not just a framework of deterrence and punishment, but also the coherence of a concept of sovereignty and non-intervention. The existing system may have functioned honourably when the only way to release water from a dam, infiltrate a federal reserve or immobilize a state was through the force of destruction-based machinery and personnel-based military operations. However, where all of these actions can be undertaken without armed force and prepared years prior in anticipation of future conflict, traditional IHL and IHRL frameworks no longer provide a constraint on gross militaristic and adversarial action.

III. A SOLUTION: THE MARTENS CLAUSE?

As foreshadowed, it is clear that some aspects of IHL can spill the banks of the peace/war paradigm.[49] A wider literature highlights that most discussions around the Martens Clause currently regard it as a creature of IHL and thus confined to its application.[50] Under this view, once interactions fail to meet the threshold of armed conflict, the Clause in any of its forms would be inapplicable. As has also been discussed, armed conflict appears comparatively comfortable in its regulation of cyber actions such as data poisoning once conflict exists, meaning the real issue remains the pre-conflict regulation of the upstream actions. This paper's view of the Clause's applicability outside of IHL is founded on Additional Protocol I and its wording.

For the benefit of contextualising the position, Article 3 of Additional Protocol I states, "the Conventions and this Protocol shall apply from the beginning of any situation referred to in Article 1 of this Protocol".[51] Article 1(3) states, "[t]his Protocol, which supplements the Geneva Conventions of 12 August 1949 for the

---

[49] Christopher Greenwood, "The Concept of War in Modern International Law", *International and Comparative Law Quarterly,* Vol. 36, No. 4, 1987, p. 283.

[50] Israel, "Comments from the State of Israel on the International Law Commission's Draft Principles on the Protection of the Environment in Relation to Armed Conflicts as Adopted by the Commission in 2019 on First Reading" (Comments Submitted to the International Law Commission), 2020, para. 56, available at: http://legal.un.org/docs/?path =../ilc/sessions/73/pdfs/english/poe_israel.pdf&lang=E; United Kingdom, "Comments and Observations on the Draft Principles on Protection of the Environment in Relation to Armed Conflicts" (Comments Submitted to the International Law Commission), 2021, paras. 21–23, available at: http://legal.un.org/docs/?path=../ilc/sessions/73/pdfs/english/ poe_uk.pdf&lang=E; United States, "Comments of the United States on the International Law Commission's Draft Principles on the Protection of the Environment in Relation to Armed Conflicts" (Comments Submitted to the International Law Commission), 6 October 2021, pp. 10–11, available at: http://legal.un.org/docs/?path=../ilc/sessions/73/pdfs/english /poe_us.pdf&lang=E; see further Dieter Fleck, "The Martens Clause and Environmental Protection in Relation to Armed Conflicts", *Goettingen Journal of International Law,* Vol. 10, No. 1, 2020, 250.

[51] AP I, Art. 3.

protection of war victims, shall apply in the situations referred to in Article 2 common to those Conventions."[52] Common Article 2 of the Geneva Conventions states, "[i]n addition to the provisions which shall be implemented in peacetime, the present Convention shall apply to all cases of declared war or of any other armed conflict which may arise between two or more of the High Contracting Parties, even if the state of war is not recognized by one of them."[53]

Here, this paper focuses on the statement "[i]n addition to the provisions which shall be implemented in peacetime".[54] Whilst it is recognized this refers generally to articles on dissemination of information and restriction of the use of emblems, this paper argues that Article 1 of Additional Protocol I, read with Common Article 2 of the Geneva Conventions, includes itself in such peacetime application through its direct and intentionally broad identification of "cases not covered by this Protocol",[55] as well as through Additional Protocol I's requirement that "[t]he High Contracting Parties undertake to respect and to ensure respect for this Protocol in all circumstances."[56] As such, the Clause may remain active in peacetime through a combination of this ambiguity and the explicit application to "cases not covered".[57] Not only this, but it is clearly apparent that the contemporary fulfilment of the Clause's purpose both requires and actively dictates its application externally to the confines of traditional IHL.

Additional Protocol I complements the Geneva Conventions and sits within their broader compliance architecture. Common Article 2 expressly acknowledges that some provisions operate "in peacetime", which supports the idea that the law of armed conflict is partly operationalized in advance through standing duties and anticipatory compliance disciplines.[58] The sentiments of the Clause are further removed from the confines of general provisions and principles in IHL treaty architecture through the articles on denunciation in each of the four Geneva Conventions. Article 63 of Convention I, Article 62 of Convention II, Article 142 of Convention III and Article 158 of Convention IV all state:

> The denunciation shall have effect only in respect of the denouncing Power. It shall in no way impair the obligations which the Parties to the conflict shall remain bound to fulfil by virtue of the principles of the law of nations, as they result from the usages established among civilized

[52] *Ibid.,* Art. 1(3).

[53] Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 12 August 1949, 75 UNTS 31 (entered into force 21 October 1950), Art. 2.

[54] *Ibid.*

[55] AP I, Art. 1(2).

[56] Ibid., Art. 1(1).

[57] Ibid., Art. 1(2).

[58] GC I, Art. 2.

> peoples, from the laws of humanity and the dictates of the public conscience.[59]

Whilst this speaks directly to parties to a conflict, it shows that the obligation to the principles of humanity and public conscience continues to exist outside the parameters of the Conventions through their endurance of any denunciation. Furthermore, at no point do these articles limit those obligations to any conflict currently ongoing.

In *Corfu Channel (UK v. Albania)*, the ICJ when explaining Albania's legal duty to warn British ships of the mines in its waters, found that among other things, the duty arose from an obligation to "elementary considerations of humanity" which the Court described as "even more exacting in peace than in war". [60] This demonstrated international law's recognition of similar humanity-based obligations existing outside armed conflict.

Additionally, as discussed, the ICJ in the Nuclear Weapons Advisory Opinion illustrated, that Additional Protocol I when adopted, acted to represent pre-existing customary law including the Clause and that the absence of codified law in regard to a form of weaponry being detailed in existing "does not permit the drawing of any legal conclusions relating to the substantive issues which the use of such weapons would raise."[61] In addition to this, and in illustration of the way in which capabilities such as those possessed by the Clause are both able and obliged to police the gaps in law, the Court also stated "humanitarian law, at a very early stage, prohibited certain types of weapons either because of their indiscriminate effect on combatants and civilians or because of the unnecessary suffering caused to combatants, that is to say, a harm greater than that unavoidable to achieve legitimate military objectives. If an envisaged use of weapons would not meet the requirements of humanitarian law, a threat to engage in such use would also be contrary to that law."[62]

In explanation of the decision, President Bedjaoui stated that "[i]nternational humanitarian law is a particularly exacting corpus of rules, and these

---

[59] GC I, Art. 63; Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea of 12 August 1949, 75 UNTS 85 (entered into force 21 October 1950), Art. 62; Geneva Convention (III) relative to the Treatment of Prisoners of War of 12 August 1949, 75 UNTS 135 (entered into force 21 October 1950), Art. 142; Geneva Convention (IV) relative to the Protection of Civilian Persons in Time of War of 12 August 1949, 75 UNTS 287 (entered into force 21 October 1950), Art. 158.

[60] International Court of Justice, *Corfu Channel (United Kingdom v. Albania),* Judgement (Merits), *ICJ Reports 1949*, p. 22.

[61] *Legality of the Threat or Use of Nuclear Weapons*, above note 13, p. 259, para. 84.

[62] *Legality of the Threat or Use of Nuclear Weapons,* above note 13, p. 257, para. 74.

rules are meant to be applied in all circumstances."[63] Here the emphasis is that IHL's obligations are not loose and cannot be suspended by a particular set of circumstances. This paper suggests that the same view applies in evaluation of peacetime. Just because armed conflict is not present does not excuse a state from compliance with the humanitarian obligations of IHL, specifically those achievable through utilisation of the Clause.

The next foreseeable opposition to the application of the Clause in the proposed format is the general uncertainty over its function and ability. In pre-emptive response, it is easiest to identify what the Clause does not apply to: anything already codified. This foundational principle of the Clause is uncontroversial and illustrates it as a backstop operating where codified coverage runs out. With this view, the Clause functions like a net that catches what slips through the confines of existing law. However, that mandate is not stationary in time. As echoed in the views of the ICJ, the Clause's real significance and likely prominence in a world where treaty-making continues to slow is its ability to operate when technology and capabilities evolve faster than codified law can be developed.[64]

A further step in sketching the application of the Clause is to explain that its wording and implied mandate do not only call it to action when the stakes are high; rather it, exists as an ongoing check on legality sitting at the boundary of what is yet to be codified.

If there are no codified rules as to a particular technology, action, or capability, then the existence of the Clause illustrates that this does not, by virtue of absence, make the technology, action, or capability legally permitted. Here, the Clause's existence sits in direct opposition to the *Lotus* principle and *a contrario* thinking.[65] This principle, taking its name from *The Case of the S.S. 'Lotus'*, decided by the Permanent Court of International Justice in 1927, is commonly used to describe the view that States may act unless international law expressly prohibits the conduct in question.[66] Accordingly, the Clause functions in diametric opposition, implying that conduct is not presumptively permitted through a mere lack of express prohibition.

Illegality without formal codified prohibition of conduct, is not a foreign concept. In the *K.W.* case,[67] the defendant was accused of intentionally injuring

---

[63] Ibid., Declaration of President Bedjaoui, p. 272, para. 19.

[64] Ibid., p. 257, para. 78.

[65] Michael Salter, "Reinterpreting Competing Interpretations of the Scope and Potential of the Martens Clause", *Journal of Conflict and Security Law,* Vol. 17, No. 3, 2012, p. 403.

[66] Permanent Court of International Justice, *The Case of the S.S. 'Lotus' (France v. Turkey)*, Judgment, 7 September 1927, *PCIJ Reports,* Series A, No. 10.

[67] Conseil de guerre de Bruxelles (Military Court of Brussels), *K.W.,* Opinion and Judgment, 8 February 1950, *Revue de droit pénal et de criminologie,* Vol. 30, No. 5, 1949–50, p. 562.

civilians who were resistant to the German occupation of Belgium. During the proceedings, the Court noted the defendant's alleged actions were not expressly forbidden by the Hague Regulations,[68] instead finding that through utilisation of the Clause, principles of international law such as the *Universal Declaration of Human Rights* could still be applied to illustrate the illegality of the actions. Through this, the Court ruled the defendant had breached Art. 5 of the Universal Declaration and thus his actions constituted a violation of the customs of war.[69]

For the Clause to operate in this way, it is the responsibility of the state to evaluate and justify the proposed action's compliance with the principles of humanity and the dictates of public conscience. This embodiment of the Clause does not require a baseline trigger beyond a case not being covered by existing international law, as it is supported by its wording and the mandate which Additional Protocol I gives it to remain active in evaluating everything untouched by codified law. The Clause creates a presumption in favour of protection. It is the duty of the state seeking to take action not yet governed by codified law to ensure and justify that such action does not breach these thresholds.

For the Clause to act as proposed above, it requires a foundation to be supplied for the baseline meanings or practical invocation of the principles of humanity and dictates of public conscience. For this, the best available illustrations come from international judicial decisions.

In *Kupreškić,* the Court did not attempt to claim that individually the principles of humanity and dictates of public conscience, had "been elevated to the rank of independent sources of international law".[70] Instead, it showed that the Clause requires, at minimum, recourse to the principles of humanity and dictates of public conscience where a rule of IHL is not sufficiently rigorous or precise, so that the scope of the rule can be defined by reference to them.

This demonstrated that whilst the Court treated the principles of humanity and dictates of public conscience as being anchored in pre-existing law, their operation was not confined to the strict limits of codified rules. Here, a customary or contemporary position may utilize the Clause and relevant considerations to constrain action that is *prima facie* non-restricted when viewed only in comparison to existing law. The Court continued to illustrate the Clause's capacity to interact with

---

[68] Antonio Cassese, "The Martens Clause: Half a Loaf or Simply Pie in the Sky?", *European Journal of International Law,* Vol. 11, No. 1, 2000, p. 207.

[69] Marcus Hickleton, "Shaky Foundations: Killer Robots and the Martens Clause", *Perth International Law Journal,* Vol. 4, 2019, p. 52.

[70] International Criminal Tribunal for the Former Yugoslavia, *Prosecutor v. Kupreškić et al.,* Case No. IT-95-16-T, Judgment (Trial Chamber), 14 January 2000, p. 206, para. 525 [hereinafter *Kupreškić*].

various forms of international law, showing how it can still define boundaries even when hard law or consistent state practice is lacking.[71]

The Court found that use of the Clause does not require evident state behaviour in the form of consistent state practice. Instead, identifying that States' belief about what the law requires may be more important than the actions they take, and that binding law can emerge and develop from such a process.[72] Further to this, the court did not treat public conscience as an incalculable moral idea. It demonstrated how *opinio necessitatis* can be evidenced in practice through military manuals, ratification patterns, ICRC memos and UN General Assembly resolutions.[73] The court used these *opinio necessitatis* identifiers as indicators of public conscience to support its implementation of the Clause. The court did not survey or attempt to codify global opinions. Rather, it evaluated how law applies, and how it ought to apply. This showed that the principles of humanity and dictates of public conscience are not freestanding morality-based vetos but are instead indicative criteria used to gauge the acceptability of the effects and interactions with humanitarian protections of actions which are not governed by current law.

In understanding principles of humanity, the identification of these principles should not be confined to IHL or IHRL alone, as the conduct being assessed will not yet be codified and contained within any single international legal apparatus; it would be an absurd notion to restrict the application of humanitarian principles to their required identification with any arbitrary legal paradigm. From this, principles of humanity should be drawn from existing law, custom, and scholarship structured to protect humanitarian interests across a broad range of international legal structures. To limit the Clause's applicability in this respect to only certain applications or structures would again, negate the breadth of its wording and mandate. As Jonathan Kwik has noted:

> One can easily imagine the backlash and media attention generated from a perimeter defence turret being briefly fooled into shooting defenceless villagers, because an adversary handed out innocuous bags at the market with adversarial patterns on them.[74]

This is exactly the kind of "moral outrage" that makes the most sense to channel via the Martens Clause. Even if lawyers could construct technical arguments that such conduct falls outside existing prohibitions, the instinctive reaction of most societies would remain the same: the act is morally repugnant. That reaction matters. The Martens Clause exists precisely to capture this category of conduct, where emerging technologies outpace codified rules but nevertheless offend fundamental

---

[71] *Ibid.,* p. 207, para. 527.

[72] *Ibid.*

[73] *Ibid.,* p. 210, para. 532.

[74] J. Kwik, above note 3, p. 744.

expectations of humanity and conscience. In this respect, the Clause is not a vague moral appeal, but a mechanism designed to ensure that legal restraint survives even when adversaries weaponize ambiguity itself.

From this, a combination of existing IHL, IHRL and customary positions should facilitate the utilisation of the principles of humanity in the Clause's application. As such, in illustration of its guiding principles, to anchor any risk-based assessment, the International Covenant on Civil and Political Rights states that "[e]very human being has the inherent right to life. This right shall be protected by law. No one shall be arbitrarily deprived of his life."[75] The facilitation of this right expands to an obligation where the threshold of the requirement for a state to act is met by the possibility of a loss to life not just the eventuation of a loss of life.[76] Such an obligation to ensure facilitation of this right is also engaged by a threat to infrastructure, social requirements and civilian systems which when impeded or restricted would pose a foreseeable threat to survival.[77] Additional to this and in argument for the adequate protection of this right in the contemporary world, it would be paradoxical to confine the civilian's "general protection against dangers arising from military operations" provided by Additional Protocol I Article 51 to scenarios of armed conflict.[78]

IV. CONCLUSION

Pre-conflict data poisoning exposes a structural weakness in international law. Existing legal frameworks remain heavily dependent upon a binary distinction between peace and war, yet contemporary cyber operations are deliberately designed to exploit the space between them. Data poisoning does not require kinetic force, immediate destruction, or even visible disruption to produce strategic effect. Instead, it enables adversaries to quietly manipulate the informational environment years before conflict emerges, embedding vulnerabilities into the systems upon which militaries, governments, and civilian societies increasingly depend. In doing so, these operations undermine the coherence of legal regimes built around temporally identifiable moments of armed conflict.

It is within a gap of strict interpretation of the two paradigms of international law, that the Martens Clause becomes legally and strategically significant. Its historical function has been to ensure that technological innovation does not outpace legal restraint, as illustrated in the ICJ's Nuclear Weapons Advisory Opinion and in

---

[75] International Covenant on Civil and Political Rights, 999 UNTS 171, 16 December 1966 (entered into force 23 March 1976), Art. 6(1).

[76] Human Rights Committee, General Comment No. 36: Article 6: Right to Life, UN Doc. CCPR/C/GC/36, 30 October 2018, p. 2, para. 7.

[77] *Ibid.,* p. 6, para. 26.

[78] AP I, Art. 51(1).

subsequent jurisprudence.[79] It clearly **exists**, but the nature of its relevance is the matter for debate. It is the opinion of the authors that, if a State merely regards the Martens Clause as a principle (rather than a legal obligation), then it is not effective. It actually risks weakening the structure of IHL, and furthering the divide between those that do, and those that think.

Far from a symbolic relic of humanitarian idealism, the Clause exists precisely to govern situations not yet contemplated by codified law. Whilst applying the Clause does not create bright-line rules or resolve attributional or enforcement issues, applying it in a manner consistent with the preamble to the 1899 Hague Convention, Additional Protocol I and decisions of the ICJ, allow it to provide a principled basis for restraint on action both prior to and during the existence of armed conflict.[80] By invoking the principles of humanity and dictates of public conscience, the Clause provides a framework capable of evaluating emerging technologies whose effects transcend traditional categories of peace, war, civilian infrastructure, and military capability. In this respect, data poisoning resembles previous moments of technological rupture, where law struggled to keep pace with innovation, but humanitarian principles remained operative.

Ultimately, the challenge posed by pre-conflict data poisoning is not simply one of cyber regulation, but of legal imagination. International law cannot remain credible if hostile actions capable of shaping future conflict are permitted to evade scrutiny merely because they occur before the formal outbreak of hostilities. The Martens Clause offers the most coherent existing mechanism for preventing that outcome. By applying the principles of humanity and the dictates of public conscience to the manipulation of datasets that shape automated decision-making, the Clause restores a measure of normative oversight to an otherwise unregulated domain of strategic competition. In doing so, it reaffirms a central proposition of humanitarian law: that the evolution of technology must never be allowed to outpace the protection of humanity.

---

[79] *Legality of the Threat or Use of Nuclear Weapons,* above note 13; *Kupreškić,* above note 70; A. Cassese, above note 68.

[80] Hague Convention II, above note 18, Preamble; AP I, Art. 1(2); *Legality of the Threat or Use of Nuclear Weapons,* above note 13.

PART 2
REGULATORY MODELS

Coding the Laws of War: Will Treaty-Following AI collapse into the dichotomy between 'apology' and 'utopia'?

Sze Hong Lam[*]

ABSTRACT

Conceptualising the Law of Armed Conflict/International Humanitarian Law (LOAC/IHL) as a legal 'superposition', this paper uses it as a 'hard case' to challenge three implicit assumptions of the Treaty-Following AI (TFAI) framework: genuine agreement, objective determinacy, and neutral interpretation. The paper advances 'strategic ambiguity' as the meta-indeterminacy that explains why LOAC/IHL's four established indeterminacies — linguistic, legal, structural, and practice — persist. While TFAI could improve compliance with the determinate core, its application to the contested penumbra risks collapsing the discursive space that legitimises LOAC/IHL. The paper advocates 'contestability by design', in which TFAI should flag rather than resolve legal indeterminacy.

Keywords: Artificial Intelligence; Treaty-Following AI; Alignment; Indeterminacy; International Humanitarian Law; Strategic Ambiguity

Introduction

In one of Isaac Asimov's short science-fiction stories published in 1947, "Runaround", the soon-to-be giant of modern science fiction posed a deceptively simple question: Could autonomous agents be reliably constrained by embedded legal rules?[1] The genius of Asimov's short stories lay in recognising that merely embedding legal rules into technical architecture does not eliminate moral and ethical dilemmas; instead, it ingrains and amplifies them through interpretation. Seventy-five years later, as Artificial Intelligence (AI) grows increasingly capable of

[*] Dr Sze Hong Lam is a Postdoctoral Fellow at the Centre for International Law (CIL), National University of Singapore (NUS). Email: lamsh@nus.edu.sg. ORCID: https://orcid .org/0000-0002-3481-4665. The author would like to thank Cullen O'Keefe, Janna Tay, and the teams at the Institute for Law & AI (LawAI) for organising the Workshop on Law Following AI; and Dr Samuel White, Ayma Naseem and the event teams at CIL for organising the Lieber AI Manual Workshop, where the author had the chance to present the earlier drafts of this paper. Special thanks to Dr Matthijs Maas (LawAI), Dr Jonathan Kwik (T.M.C. Asser), Professor Rebecca Crootof (Richmond), Professor Thomas Burri (HSG), Dr Talita Dias (Exeter), Professor Jon Truby (CIL), Professor Sean Watts (Lieber Institute), Professor Joseph H.H. Weiler (NYU), Dr Bethan Hall (CIL), Dr Tine Deschuytere (CIL), Dr Rishabh Bajoria (CIL), Vincenzo Elia (CIL), and Liu Mei Ching (RSIS) for engaging with the earlier drafts and providing helpful feedback. He would also like to thank the editorial teams for the helpful editing. Any errors belong solely to the author and the author alone.

[1] Issac Asimov, *I, Robot*, Gnome Press, New York, 1950, pp. 20–21.

acting autonomously across domains of strategic importance, the question posed by Asimov in 1947 has become urgently practical.[2]

In 2025, Matthijs Maas and Tobi Olasunkanmi, two researchers from the Boston-based Institute for Law & AI, jointly proposed the framework of Treaty-Following AI (TFAI), in which agentic AI systems are designed to generally follow their principals' instructions unless they entail a violation of the terms and obligations of a benchmark treaty (AI-Guiding Treaty).[3] Maas and Olasunkanmi argued that TFAI could provide States with a credible, self-executing commitment mechanism that shields the principals from legal risk arising from the AI system's unforeseen illegal behaviour in high-stakes environments, including armed conflicts.[4] This proposal is not entirely fictional, as Scale AI's *Defence Llama*, an operation planning AI system, has already been trained on the US Department of Defence's Guidelines for Armed Conflicts and Ethical Principles for AI (DoD Guidelines).[5]

In light of the *de facto* improbability of signing a tailor-made treaty regulating the use of military AI,[6] one likely way to implement the proposal for military agentic AI would be to turn the four Geneva Conventions of 1949[7] and the two Additional Protocols of 1977[8] into AI-Guiding Treaties.[9] To visualise the operation of TFAI,

---

[2] As early as in 2009, Ronald Arkin has already suggested the idea of 'moral-rule-sensitive-robots', and its potential deployment in military context, see generally, Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, Boca Raton, 2009; the idea of moral robot also appeared in another research sponsored by the US Office of Naval Research in 2008, see Patrick Lin, George Bekey and Keith Abney, *Autonomous Military Robotics: Risk, Ethics, and Design*, Ethics + Emerging Sciences Group at California Polytechnic State University, San Luis Obispo, December 2008.

[3] Matthijs M. Maas and Tobi Olasunkanmi, *Treaty-Following AI*, Institute for Law & AI Working Paper No.1 of 2025, available at: https://law-ai.org/treaty-following-ai/ (all internet references were accessed in May 2026).

[4] *Ibid*, p.12.

[5] Scale AI, "Introducing Defense Llama", 5 November 2024, available at: https://scale.com /blog/defense-llama.

[6] See, e.g., Zhixiong Huang and Haokun Yu, "*De Jure* Necessity vs *De Facto* Improbability? Establishing a New Treaty for Artificial Intelligence in the Military Domain", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.129-159.

[7] Geneva Convention (I) for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 12 August 1949, 75 UNTS 31 (entered into force 21 October 1950); Geneva Convention (II) for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea of 12 August 1949, 75 UNTS 85 (entered into force 21 October 1950); Geneva Convention (III) relative to the Treatment of Prisoners of War of 12 August 1949, 75 UNTS 135 (entered into force 21 October 1950); Geneva Convention (IV) relative to the Protection of Civilian Persons in Time of War of 12 August 1949, 75 UNTS 287 (entered into force 21 October 1950).

[8] Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978); Protocol Additional (II) to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts, 1125 UNTS 609, 8 June 1977 (entered into force 7 December 1978).

[9] For similar ideas, see, e.g., Mark Chinen, "The Co-Evolution of Autonomous Machines and Legal Responsibility", *Virginia Journal of Law & Technology*, Vol. 20, No. 2, 2016, p .343; & Scott Sulivan, who

imagine an agentic AI system that receives an order from a human commander to plan a drone strike. Before turning the instructions into targeting orders, the TFAI framework requires the AI to run the following compliance test: (1) determine whether the goal is *prima facie* illegal (e.g. attacking only civilians), in which case it refuses to execute; (2) otherwise, identify likely treaty issues such as precaution, distinction, and proportionality, and reason through possible interpretations; (3) where the treaty text leaves room for ambiguity, consult supplementary sources; (4) evaluate whether the resulting legal risk falls within the 'acceptable' range defined by the AI-Guiding Treaty; (5) if so, relay orders to the drones; if not, refuse to execute or provide alternative action plans (the "hypothetical scenario").[10] This compliance test instantiates an *ex ante* model of legal application, which the rest of the paper interrogates.

The TFAI proposal itself acknowledges that it offers only a minimum viable use case for a novel commitment mechanism and that extension of TFAI to all areas of international law is likely not its goal. This paper takes this acknowledgement seriously by using the Law of Armed Conflict/International Humanitarian Law (LOAC/IHL), a discipline that stands at 'the vanishing point of international law',[11] to test TFAI's boundary conditions.[12] This paper identifies three domain-specific characteristics of LOAC/IHL that place significant pressure on the applicability of TFAI: (1) genuine agreement; (2) objective determinacy; and (3) neutral interpretation. Defining "determinacy" as a situation when law only gives one correct answer, and "indeterminacy" as when there is more than one legally defensible answer.[13] This paper draws on existing theories of legal indeterminacy in international law. It argues that the four established indeterminacies of LOAC/IHL — linguistic, legal, structural, and practice — are sustained by a fifth and more fundamental source of indeterminacy: "strategic ambiguity", which this paper

---

linked IHL-by-design to weapons review under Article 36 of the 1977 First Additional Protocol, see Institute for Law & AI, *Law-Following AI in International Humanitarian Law | Workshop on Law-Following AI 2025* (2026), available at: https:// www.youtube.com/watch?v=Dy07B2pBGZE.

[10] This scenario is built on the steps identified in M. Maas and T. Olasunkanmi, above note 3, p. 32; Tomasz Zurek, Jonathan Kwik, and Tom van Engers have already constructed a computational framework based on the NATO Standard Targeting Procedure. See, generally Tomasz Zurek, Jonathan Kwik and Tom Van Engers, "Model of a Military Autonomous Device Following International Humanitarian Law", *Ethics and Information Technology*, Vol. 25, No. 15, 2023, p. 15.

[11] 'If international law is, in some ways, at the vanishing point of law, the law of war is, perhaps even more conspicuously, at the vanishing point of international law.' See, Hersch Lauterpacht, "The Problem of the Revision of the Law of War", *British Yearbook of International Law*, Vol. 29, 1952, p. 382.

[12] Critical scholarship has examined the prospects of automation in contract law, regulatory law, and adjudication […]' See, e.g., Simon Deakin FBA and Christopher Markou (eds), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence*, Hart Publishing, an imprint of Bloomsbury Publishing, Oxford, 2020. It has not yet addressed the specific challenges posed by LOAC/IHL, a domain whose constitutive ambiguities, this paper argues, place it at the far end of the spectrum of computability.

[13] Timothy Endicott, *Vagueness in Law*, Oxford University Press, Oxford, 2003, p. 9; This is equivalent to what the field of AI & Law has defined as a condition for 'hard' legal questions. See, e.g., Kevin D. Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*, Cambridge University Press, Cambridge, 2017, p. 20.

advances as the "meta-indeterminacy" that explains why the others persist. Together, they test the three TFAI assumptions to their limits.

Borrowing loosely from quantum mechanics,[14] this paper uses the term LOAC/IHL to denote a "superposition" of the same body of law that regulates the conduct of participants during an armed conflict (*jus in bello*).[15] Here, this paper adopts the same definition of the two acronyms used by David Luban in 2012: Law of Armed Conflict (LOAC), which is commonly used by military lawyers and States' legal advisors to highlight the importance of military advantage; and International Humanitarian Law (IHL), preferred by human rights lawyers, academics, and non-governmental organisations (NGOs) to emphasise the humanitarian vision.[16] This metaphor captures three features of LOAC/IHL simultaneously: (1) like light that is both wave and particle, LOAC/IHL holds two logically irreconcilable visions at once; (2) like a photon existing in multiple states until observed, LOAC/IHL's key principles remain indeterminate prior to application; and (3) just as observation collapses a superposition, hard-coding reduces LOAC/IHL's indeterminacy to a singular vision of law-regulating warfare.[17] The metaphor highlights LOAC/IHL's interpretive plurality, suspended by its very indeterminacy, and its desirable role in reconciling two irreconcilable visions of the same body of law.

This paper's core claim is that indeterminacies in LOAC/IHL were not merely tolerated by their drafters but actively preserved, and that States have continuing interests in maintaining rather than collapsing this superposition. Yet even within these boundary conditions, TFAI offers genuine promise: it can reduce LOAC/IHL violations rooted not in legal ambiguity but in the human and organisational factors, such as fear, retaliation, fatigue, and peer pressure, that have historically accounted for most violations of the law. However, TFAI's potential is limited precisely where LOAC/IHL rules are "strategically ambiguous". This paper is structured into three different parts: Section 1 details TFAI and its assumptions. Section 2 uses various indeterminacy theories to map out the extent of LOAC/IHL's determinate core versus its contested penumbra and identify the sources of the indeterminacies before considering whether they could be 'cured' by

---

[14] For a similar analogy in other branches of international law, see generally, e.g., Xinyue Li, *International Economic Law and Quantum Theory: Quantising Geoeconomics*, Routledge, London, 2026.

[15] While the Law of Armed Conflict (LOAC) is sometimes used more broadly to also include the body of international law that regulates when it is lawful to be involved in an armed conflict (*jus ad bellum*), this paper defined LOAC in a more narrow sense as only relating to the law regulating conduct of hostilities (*jus in bello*). For a critique of this dual conceptualisation, see, e.g., Carsten Stahn, ""*Jus Ad Bellum*", "*jus in Bello*" . . . "*Jus Post Bellum*"? -Rethinking the Conception of the Law of Armed Force", *European Journal of International Law*, Vol. 17, No. 5, 2006, pp. 925-926.

[16] David Luban, "Military Necessity and the Cultures of Military Law", *Leiden Journal of International Law* Vol. 26, No. 1, 2013, p. 316.

[17] This paper adopts the 'Copenhagen interpretation' of Niels Bohr and Werner Heisenberg to understand quantum mechanics. The same school of interpretation is regarded as synonymous with indeterminism. See, e.g., Jan Faye, "Copenhagen Interpretation of Quantum Mechanics", in Edward N. Zalta and Uri Nodelman (eds), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University, Stanford, 2024, available at: https://plato.stanford.edu/archives/sum2024/entries/qm-copenhagen/.

technical fixes. Instead of seeking a single, stable, objective rule that can be "followed" by the AI system, section 3 of this paper advocates a "contestability by design" feature that recognises TFAI's potential to improve compliance with the determinate core of LOAC/IHL while preserving the role of human lawyers in the interpretation of LOAC/IHL in military decisions.

1. Treaty-Following AI: The three fundamental assumptions

1. Genuine Agreement

Regarding TFAI, Maas and Olasunkanmi stated that their framework was modest. Instead of contributing to human compliance with international law, TFAI is termed "technical *ex ante* infrastructural constraints", in which treaty text is treated as "an appropriate, stable, and certified referent text".[18] As such, Maas and Olasunkanmi compared TFAI to existing AI industry safety techniques, such as Anthropic's "Claude constitution"[19] and OpenAI's Model Spec.[20] The TFAI proposal must also be considered alongside a growing body of scholarship calling for the alignment of military AI with customary LOAC/IHL principles.[21] Maas and Olasunkanmi highlighted a particular risk associated with an "AI henchman", who faithfully implemented the instructions of the principal at the expense of breaching peremptory norms,[22] including the "basic rules of IHL".[23] Compared with crimes against humanity or aggression, the treaty basis of these basic rules provides TFAI with a more legitimate anchor 'under a positivist and consent-based account of international law' for legal alignment.[24]

A core part of the TFAI proposal was built on the concept of the AI-Guiding Treaty.[25] The TFAI proposal's emphasis on the treaty as a legitimacy anchor drew on multiple sources. The argument for the "greater legitimacy" based on the State's consent to treaty was partially inspired by a mirror proposal of Law-Following-AI

---

[18] M. Maas and T. Olasunkanmi, above note 3, p. 60.
[19] Yuntao Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback" *arXiv*, 2022, available at: https://doi.org/10.48550/ARXIV.2212.08073.
[20] "OpenAI Model Spec", available at: https://model-spec.openai.com/2025-02-12.html.
[21] See, e.g., Ashley Deeks, "Coding the Law of Armed Conflict: First Steps", in Matthew C. Waxman and Thomas W. Oakley (eds), *The Future Law of Armed Conflict*, 1st ed., Oxford University Press, New York, 2022; John Mavi, Diana Teodora Găitan and Sergio Coronado, "From Rogue to Safe AI: The Role of Explicit Refusals in Aligning LLMs with International Humanitarian Law", *arXiv*, 5 June 2025, available at: https://doi.org/10.48550/arXiv.2506 .06391.
[22] M. Maas and T. Olasunkanmi, above note 3, p. 19.
[23] See, conclusion 23 Annex (d) in International Law Commission, "Draft Conclusions on Identification and Legal Consequences of Peremptory Norms of General International Law (Jus Cogens), with Commentaries", UN Doc. A/77/10, available at: https://legal.un.org/ilc/texts/instruments/english/commentaries/1_14_2022.pdf.
[24] Institute for Law & AI, *Treaty-Following AI | Workshop on Law-Following AI 2025*, 12:00, available at: https://www.youtube.com/watch?v=huORIiQNB54v.
[25] M. Maas and T. Olasunkanmi, above note 3, p. 60.

(LFAI), which involves "aligning" the domestic use of agentic AI systems with core provisions of criminal law and constitutional law that are deemed sufficiently stable and democratically legitimate.[26] The TFAI proposal was also notably inspired by Bajgar and Horenovsky's earlier proposal of using 'negative human rights' to constrain AI behaviours.[27] These works fell under the relevant discussion on the benchmark to which AI systems should align, famously raised by Iason Gabriel in 2020,[28] who argued that international law and the doctrine of universal human rights should constitute a global, overlapping consensus for AI alignment.[29]

The TFAI proposal does not claim to work under all conditions or in all domains of international law. Rather, it identifies circumstances in which a specific AI-Guiding Treaty can serve as an effective benchmark. While these conditions are not always explicitly enumerated, they are implicit in the proposal's design and scope limitations. For instance, by requiring the AI-Guiding Treaty to be a "certified referent text through which States can establish jointly agreed-upon infrastructural constraints", the TFAI proposal assumed that the treaty reflected a thick consent (i.e. the real meeting of the mind between States characterised by the absence of sustained opposition to substantial issues), instead of a thin consent (i.e. the formal validation of the treaty by States).[30] This emphasis on requiring "thick consent" could be seen from the TFAI proposal, which stated the goal of TFAI was to "help crystallise an *ex ante* reciprocal commitment amongst the Contracting States, allowing them to reassure each other that they both intend to respect the intent of the treaty, not just its letter"[31] (the 'genuine agreement' thesis).

2. Objective Determinacy

Maas and Olasunkanmi established the technical feasibility of TFAI by drawing on recent research showing that AI systems have become increasingly proficient in legal reasoning, interpretation, and the application of legal norms to new cases. They highlighted how current AI systems, including the non-frontier ones, are already capable of conducting treaty interpretations that "frequently are, or appear to be, if not flawless, then nonetheless coherent, correct, and compelling to experienced international legal scholars or judges."[32] With technical feasibility established, Maas

---

[26] Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Cristoph Winter, "Law-Following AI: Designing AI Agents to Obey Human Laws", *Fordham Law Review*, Vol. 94, No. 1, 2025, p. 57.
[27] Ondrej Bajgar and Jan Horenovsky, "Negative Human Rights as a Basis for Long-Term AI Safety and Regulation", *Journal of Artificial Intelligence Research*, Vol. 76, 2023, pp. 1043-1075, available at: https://doi.org/10.1613/jair.1.14020.
[28] Iason Gabriel, "Artificial Intelligence, Values, and Alignment", *Minds and Machines,* Vol. 30, 2020, p. 411.
[29] *Ibid*., p. 427.
[30] Joost Pauwelyn, Ramses A. Wessel and Jan Wouters, "When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking", *European Journal of International Law*, Vol. 25, 2014, p. 733.
[31] M. Maas and T. Olasunkanmi, above note 3, p. 29.
[32] *Ibid*., pp. 37–38.

and Olasunkanmi proposed that the most straightforward way to operationalise TFAI is to integrate it into the AI's inference-time chain of thought.[33]

This paper does not dispute the technical feasibility of TFAI to "crack" the legal reasoning. Instead, it examines the legal feasibility of turning the Geneva Conventions and the two Additional Protocols into AI Guiding Treaties. In terms of treaty interpretation, it must be noted that TFAI's interpretation is functional. AI systems are meant to consult the treaty text only to assess the legality of potential goals or actions, not because they are themselves bound by the treaty's obligations.[34] In other words, TFAI does not need to become Ronald Dworkin's Judge Hercules, capable of finding the "right answer" in every case.[35] It needs only the "stable enough" answers that guide the AI in the same way as LOAC/IHL guides a "more-or-less" law-abiding soldier (i.e. the "objective determinacy" thesis).[36]

3. Neutral Interpretation

As stated, the TFAI proposal was inspired by the LFAI proposal, and both proposals acknowledged a series of technical constraints and challenges, including legal indeterminacy.[37] Yet, both TFAI and LFAI proponents believed that some forms of law-alignment would still be preferable to relying entirely on subjective intent-alignment and ideological-embedded value-alignment.[38] This is because it is "much easier (and less controversial) to evaluate the validity of legal propositions and arguments than to assess the quality or correctness of ethical reasoning."[39] In other words, while TFAI/LFAI is not value-neutral, it is more value-neutral than an unconstrained alignment based purely on military intent or the developer's subjective ethics. It must be noted that the TFAI paper does not claim that legal interpretation (or the resort to a TFAI framework) is value-free, but as stated before, consented-to treaty law provides a more legitimate anchor than normative codes developed by many alternative candidates, such as private AI companies, NGOs, or a single State in isolation (i.e., the 'neutral interpretation' thesis).[40]

---

[33] *Ibid*., p. 32.
[34] *Ibid*., p. 61.
[35] Ronald Dworkin, *Taking Rights Seriously*, New York, Bloomsbury Publishing, 2013, pp.38–39; This is contrary to H.L.A. Hart, who believed that such a 'correct answer' does not exist, see Geoffrey C Shaw, "H.L.A. Hart's Lost Essay: Discretion and the Legal Process School", *Harvard Law Review*, Vol. 127, 2013, p. 700.
[36] However, what is considered 'good enough' is itself a deeply political-philosophical question. See Johan Rochel, "Encoded Normativity: Designing Large Language Models for International Law", *Nordic Journal of International Law*, 2026, p. 23. Thus, a better alternative is to adopt a dialogic approach reflected in the 'panorama design' mentioned in section 3 of this paper.
[37] M. Maas and T. Olasunkanmi, above note 3, pp. 45–54; O'Keefe and others, above note 26, pp. 110–111.
[38] O'Keefe and others, above note 26, pp. 112–113.
[39] *Ibid*., pp. 115–116.
[40] M. Maas and T. Olasunkanmi, above note 3, p. 30.

In this regard, the TFAI implicitly portrays law alignment and intent/value alignment as competing or mutually exclusive. This assumption was criticised by Katalina H. Delgado for failing to account for legal pluralism, lagging law, and the letter-spirit distinction, which collapses LFAI/TFAI back into value/intent alignment.[41] This paper echoes many of Delgado's concerns. The core argument of this paper is that a relatively neutral interpretation only exists in areas of international law (1) where the objectives of the law are not mutually contradictory; (2) where there is a somewhat unified and homogeneous interpretive community; and (3) where States are not incentivised to keep the interpretation ambiguous. Indeed, if international lawyers and States broadly agree on what a treaty requires, then aligning AI systems with that shared understanding seems more legitimate than aligning them with subjective ethical views or narrow organisational interests. As the following sections of this paper will consistently show, in resolving legal indeterminacy, the choice of the interpreter ultimately hinges on how one addresses moral and ethical dilemmas that have long troubled human beings, such as the choice between saving the lives of fellow soldiers and those of enemy civilians.[42] In this sense, the legitimacy gained from treaty alignment may be indistinguishable from the existing alternatives.

2. Promises and challenges faced by the TFAI in military deployment

This paper argues that the achievability of *ex ante* consensus ultimately turns on one's view of the degree and sources of indeterminacy in LOAC/IHL. Maas and Olasunkanmi acknowledged that achieving TFAI would depend on both a technical component (TFAI agents) and a legal one (AI-guiding treaties).[43] This paper aims to precisely interrogate the second component by challenging the three identified assumptions: (1) the "genuine agreement" thesis; (2) the "objective determinacy" thesis; and (3) the "neutral interpretation" thesis, with reference to indeterminacy theories of international law, with a particular focus on LOAC/IHL.

It is uncontroversial that the degree of indeterminacy affects the feasibility of automated legal processes[44] and that public international law is, to some extent, indeterminate.[45] What is controversial is the extent of this indeterminacy both across and within the disciplines.[46] Using Severin Meier's four categories of indeterminacy as a basis,[47] this paper adds a fifth, which it advances as an original analytical

---

[41] Katalina Hernandez Delgado, "The Law-Following AI Framework: Legal Foundations and Technical Constraints. Legal Analogues for AI Actorship and Technical Feasibility of Law Alignment", *arXiv*, 2025, p.7, available at: https://arxiv.org/abs/2509.08009.

[42] In this regard, the TFAI faced the same problem as programming morality into AI systems, see, e.g., M. Chinen, above note 9, pp. 383–384.

[43] M. Maas and T. Olasunkanmi, above note 3, p. 30.

[44] See generally, Harry Surden, "The Variable Determinacy Thesis", *The Columbia Science and Technology Law Review*, Vol. 12, No. 1, 2011, pp. 1-91.

[45] Jan Klabbers, "The Meaning of Rules", *International Relations*, Vol. 20, No. 3, 2026, p. 298.

[46] Severin Meier, *Indeterminacy of International Law?*, Zürich, *sui generis* Verlag, 2021, p. 6.

[47] *Ibid*.

contribution, "strategic ambiguity". Unlike the four foregoing theories, which describe forms of indeterminacy, strategic ambiguity explains why these indeterminacies persist rather than being progressively resolved through state practice or judicial interpretation. As Section 2.5 will argue, the four established indeterminacies are sustained by States' active interest in their preservation. This paper argues that "strategic ambiguity" is therefore the "meta-indeterminacy" through which the others operate. This is presented at this stage as a hypothesis. The following four sections demonstrate this case by case and then examine how they collectively challenge the three implicit positivist presumptions behind the TFAI proposal.[48]

1. Linguistic indeterminacy

The first school of thought is linguistic indeterminacy, which broadly overlaps with legal positivism.[49] The core argument of this school is that law is indeterminate due to the vagueness and ambiguity inherent in language, leading to a stable core and a penumbra of uncertainty.[50] Simon Chesterman, for instance, pointed to the indeterminate nature of legal languages[51] and law's irreducibility to logical representation.[52] He further situated the debate as a fundamental disagreement between legal positivists and critical/legal realists over the nature of law.[53] If the positivist's view is accepted, the "objective determinacy" thesis would face technical difficulties only in the penumbra of LOAC/IHL.[54]

Similarly, LOAC/IHL scholars trace indeterminacies to the use of open-textured legal language. Ambiguity can be seen in key definitions of LOAC/IHL, such as "military objects",[55] "superfluous injury or unnecessary suffering",[56] or even

[48] This paper acknowledges that further comprehensive study is required to map out the exact extent of determinacy/indeterminacy in the entire field of LOAC/IHL, see section 0 of this paper.
[49] See, generally, S. Meier, above note 46, pp.27–29.
[50] HLA Hart, "Positivism and the Separation of Law and Morals", *Harvard Law Review*, Vol. 71, No. 4, 1958, p. 607; see also Hans Kelsen, *General Theory of Law & State*, London, Routledge, 2005, p. 146.
[51] Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, Cambridge, 2021, p. 230.
[52] *Ibid.*, pp. 230–231.
[53] Drawing from the famous Hart-Dworkin debate, Chesterman highlighted that neither Dworkin nor Hart believed that a correct answer in 'hard cases' could be reached by a computer system or logical deduction, see, e.g., *ibid.*, p. 232.
[54] René Provost, *International Human Rights and Humanitarian Law*, Cambridge University Press, Cambridge, 2002, p. 247.
[55] Ariel Zemach, "Indeterminacy in the Law of War: The Need for an International Advisory Regime", *Brooklyn Journal of International Law*, Vol. 43, No. 1, 2017, p. 8.
[56] On the indeterminacy of the meaning of 'superfluous injury', see, e.g., Henri Meyrowitz, "The Principle of Superfluous Injury or Unnecessary Suffering: *From the Declaration of St. Petersburg of 1868 to Additional Protocol 1 of 1977*", *International Review of the Red Cross*, Vol. 34, 1994, pp. 119–120.

"armed conflict".[57] Vagueness could further be a result of omission, such as the absence of a definition of "civilian"[58] and provisions governing belligerent reprisals in AP II.[59] At the same time, even when parties can agree on the definition and binding character of the norms, their meaning can often be determined only in concrete situations, meaning that some LOAC/IHL rules are bound to remain profoundly indeterminate until applied.[60] A classic example here is whether a media broadcasting centre[61] or a data centre constitutes a "military object".[62] Furthermore, some terms, such as "excessive" civilian casualties, "concrete and direct" military advantage,[63] or what counts as "feasible" precaution,[64] are inherently evaluative, leaving room for human discretion.[65] Here, Scott Sullivan, while not directly responding to the TFAI proposal, pointed out that the major difficulty in coding LOAC/IHL rules lies in translating "human standards" into a form readable by the AI system.[66] However, linguistic indeterminacy might be technically surmountable. As various studies have shown, Large Language Models (LLMs)[67] can satisfactorily resolve textual ambiguity through statistical predictions based on a sufficiently large dataset.[68] However, as the later section shows, linguistic indeterminacy in LOAC/IHL often results from deeper sources, such as discretion, the balance between military necessity and humanitarianism, and strategic ambiguity, that are not susceptible to any technical fix based on statistical prediction.[69]

---

[57] Adil Ahmad Haque, "Indeterminacy in the Law of Armed Conflict", *International Law Studies*, Vol. 95, 2019, pp. 121–122, 124; For a demonstration of how the vagueness of 'armed conflict' has been exploited by the US in Vietnam, and the Soviet Union in Afghanistan, see, e.g., R. Provost, above note 54, pp. 252–253.

[58] A. Haque, above note 57, pp. 135–138.

[59] See generally Frits Kalshoven, "Belligerent Reprisals Revisited", *Netherlands Yearbook of International Law*, Vol. 21, 1990, p. 43.

[60] A. Haque, above note 57, pp. 127–128.

[61] International Criminal Tribunal for the Former Yugoslavia (ICTY), "Final Report to the Prosecutor by the Committee Established to Review the NATO Bombing Campaign Against the Federal Republic of Yugoslavia", para. 55, available at: https://www.icty.org/ en/press/final-report-prosecutor-committee-established-review-nato-bombing-campaign-against-federal.

[62] See Ori Pomson, ""Objects"? The Legal Status of Computer Data under International Humanitarian Law", *Journal of Conflict and Security Law*, Vol. 28, No. 6, 2023, p. 349.

[63] AP I, Art. 51(4).

[64] *Ibid*., Art. 57(2)(a)(i).

[65] Sadjad Soltanzadeh, "Normative Reasoning and Military AI: DILEMA's Dilemmas", in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, TMC Asser Press, The Hague, 2026, pp. 39–40; cf a more 'objective' view of precautionary principle, which makes it more susceptible to automation, see, e.g., Elliot Winter, "The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law", *Journal of Conflict and Security Law*, Vol. 27, No. 1, 2022, pp. 17–19.

[66] See Institute for Law & AI, above note 24.

[67] An LLM is defined as a computational system trained on vast amounts of text data to predict the most probable next word in response to a given input, see J. Rochel, above note 36, p. 4.

[68] See, e.g., Yonathan Arbel and David Hoffman, "Generative Interpretation", *NYU Law Review*, Vol. 99, 2024, pp. 477–483; Richard McAdams and Christoph Engel, "Asking GPT for the Ordinary Meaning of Statutory Terms", *University of Illinois Journal of Law, Technology and Policy*, 2024, available at: https://chicagounbound.uchicago.edu/ journal_articles/10562.

[69] See sections 2.2 and 2.3 of this paper.

In the eyes of legal positivists, the indeterminacies in LOAC/IHL and domestic law share the same core and penumbra structure. Indeed, this paper believes that the TFAI proposal has the most promising potential with respect to some LOAC/IHL core provisions. For instance, the prohibition of summarily executing prisoners of war (POWs)[70] was articulated in the clearest language possible.[71] One could also, *inter alia*, add the rules relating to protective emblems[72] and the ban on the use of chemical and biological weapons to LOAC/IHL's determinate "core".[73] The technical feasibility of TFAI in LOAC/IHL has, to a certain extent, been demonstrated by the recent Luxembourg Tech School's experiment, which showed that frontier AI models could reject unlawful requests in 90% of determinate cases of violations of LOAC/IHL, including the obligations for combatants to distinguish themselves from enemies, the prohibition on the use of terror on civilians, and the unlawfulness of attacking medical and religious staff attached to armed forces.[74] Here, it is important to remember that many of the gravest LOAC/IHL violations in history have stemmed not from *ex ante* interpretive disagreement, but from fear, anger, retaliation, organisational culture, fatigue, or peer pressure.[75] Here, TFAI's potential lies in its capacity to prioritise compliance with LOAC/IHL over the human and organisational factors that may cloud the judgment of state and military lawyers in the midst of an armed conflict.[76] Thus, TFAI could play an important role in step 1 of the hypothetical scenario by filtering human instructions that *prima facie* violate LOAC/IHL rules. Here, the TFAI could be seen as operationalising the duty of human soldiers to refuse a "manifestly illegal

[70] See, e.g., AP, I Art. 75; AP II, Art. 4.
[71] See, Annyssa Bellal and Stuart Casey-Maslen, *The Additional Protocols to the Geneva Conventions in Context*, Oxford University Press, Oxford, 2022, pp. 28–38.
[72] GC I, Art. 7; AP I, Arts. 18, 85(f); AP II, Art. 12.
[73] Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction, 1015 UNTS 163, 10 April 1972 (entered into force 26 March 1975), Art. 1; Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction, 1974 UNTS 45, 3 September 1992 (entered into force 29 April 1997), Art. 1.
[74] John Mavi, Diana Teodora Găitan and Sergio Coronado, "From Rogue to Safe AI: The Role of Explicit Refusals in Aligning LLMs with International Humanitarian Law", *arXiv*, 2025, pp. 7–8, available at: https://doi.org/10.48550/ARXIV.2506.06391.
[75] Many LOAC/IHL violations are organisational-cultural rather than legal interpretative, see, e.g., Mark Osiel, *Obeying Orders: Atrocity, Military Discipline, and the Law of War*, Transaction Publishers, Piscataway, 2005; Christopher R. Browning, *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*, HarperCollins Publishers, New York, 2017.
[76] The current author agrees with Kevin Jon Heller in rejecting the simplistic notion that humans are inherently superior to machines in complying with LOAC/IHL, see, generally Kevin Jon Heller, "The Concept of "The Human" in the Critique of Autonomous Weapons", *Harvard National Security Journal*, Vo. 15, No. 1, 2023, p. 6; Kevin Jon Heller, "The Fetishization of "The Human" in the Critique of Autonomous Weapons", *Harvard National Security Journal Online*, 2025, pp. 1–11. However, while Heller addressed the ambiguities in the application of LOAC/IHL to real-life situations (which machines could indeed do better), he did not address the legal ambiguity in LOAC/IHL that requires human interpretation.

order" by embedding it in the design infrastructure of military agentic AI, thereby representing both a legal and technical improvement to the *status quo*.[77]

The argument of this paper is therefore not that TFAI is without value in LOAC/IHL, but that its value is concentrated in a particular zone, the determinate core, where compliance failures are non-interpretive. And it diminishes sharply as one moves into the contested penumbra where interpretive judgment is itself the locus of disagreement. Section 2.5 of this paper shows that several of the most operationally important linguistic ambiguities in LOAC/IHL, such as the undefined and evaluative threshold of 'excessive' in proportionality or the open texture of "military advantage", are not accidental oversights but "strategic ambiguity".[78] Their persistence across half a century of state practice, judicial interpretation, and scholarly debate is the first clue that linguistic indeterminacy in LOAC/IHL has a deeper origin than the open-texture of human language.

2. Legal indeterminacies

The second challenge relates to indeterminacies inherent in law itself. Closely associated with American realism and the New Haven School, the second school of scholars argues that law is, by nature, indeterminate because of the composition of legal norms.[79] According to this school of thought, there are some indeterminacies that are inherent in the nature of law and go beyond mere fuzziness of human languages: first, the sources of law;[80] second, the methods of interpretation; third, discretion.[81]

First, regarding sources of law, legal indeterminacy can arise in step 3 of the hypothetical scenario. When TFAI must resort to supplementary means of interpretation, the AI designers must first address questions about what constitutes a source of LOAC/IHL. This question is particularly relevant but controversial for the US, which is not a party to the two Additional Protocols and has an ongoing dispute with the ICRC over the customary status of some LOAC/IHL core provisions.[82] On this point, Ashley Deeks, former deputy legal adviser to the US National Security Council, suggested that coding LOAC/IHL principles has the unappreciated benefit

---

[77] This is an improvement because, under existing international criminal law, superior order can be a defence to war crimes; see, e.g., Rome Statute of the International Criminal Court, UN Doc. A/CONF.183/9, 17 July 1998 (entered into force 1 July 2002), Art. 33(2).

[78] Cf some indeterminacies that were accidentally created, such as the obligation to "ensure respect" in Common Article 1 of the four Geneva Conventions of 1949, see Boyd van Dijk, *Preparing for War: The Making of the Geneva Conventions*, Oxford University Press, Oxford, 2022, pp. 299–300.

[79] S. Meier, above note 46, pp. 37–45.

[80] In Meier's classification, indeterminacy relating to sources of law was framed as a structural critique, but as he acknowledged, neither Kennedy nor Koskenniemi would frame their claim as relating to the sources, but rather, to legal validity, see *ibid*., p. 51.

[81] *Ibid*., p. 36.

[82] See, e.g., Michael N. Schmitt, "Lieber Papers Series - The United States and (Most of) the Rest: A Legal Interoperability Primer", *Lieber Institute West Point*, 18 October 2024, available at: https://lieber.westpoint.edu/united-states-most-rest-legal-interoperability-primer/.

of forcing a government to reach consensus *ex ante* on the status of certain LOAC/IHL concepts.[83] While there is no doubt that TFAI might help forge an intra-state consensus, there is still a question of "democratic deficit" when rules governing military AI are set through negotiations between militaries and big tech companies.[84] This deficit could be exacerbated by human reinforcement learning, in which the selection of human experts and the positive feedback loops might entrench TFAI's hidden interpretive choice without external scrutiny.[85] Section 2.4 of this paper will return to the point of "inclusiveness" in AI design in light of the division between military and humanitarian lawyers.

Indeterminacy in the sources of law often leads to fragmentation of applicable rules, both within and beyond the LOAC/IHL field, through inconsistent rules and the co-application of parallel legal regimes to the same factual scenario.[86] Returning to step 2 of the hypothetical scenario, in the absence of a clear objective standard for how accurate a weapon system must be to be considered "specifically directed",[87] the AI system can assess the legality of the targeting decision either under the principle of distinction or proportionality, with each yielding a permissive or restrictive outcome.[88] In other words, a major source of indeterminacy in LOAC/IHL regarding military targeting lies in the fact that the boundary between distinction and proportionality is less clear than most LOAC/IHL lawyers would have assumed.[89] Against this indeterminacy, legal positivists would have argued that this could be easily resolved by correctly applying the secondary rules of recognition. Yet what counts as a correct application in a decentralised legal order? The indeterminacy lies exactly here.[90] One could observe this in the never-ending debates over whether LOAC/IHL should be systematically integrated with international

[83] A. Deeks, above note 21, p. 44.
[84] See, e.g., Alan Rozenshtein, "Congress—Not the Pentagon or Anthropic—Should Set Military AI Rules", *Lawfare*, 20 February 2026, available at: https://www.lawfaremedia.org/ article/congress-not-the-pentagon-or-anthropic-should-set-military-ai-rules.
[85] See, e.g., J. Rochel, above note 36, pp. 17–24.
[86] AP I, Arts. 37, 40, 44, 49(1); A. Haque, above note 57, pp. 135–138.
[87] See, e.g, the controversies relating to the 200-metre standard set in the *Gotovina et al* case in section 2.2 of this paper.
[88] For a real-life example in which the same attack was interpreted differently under the principle of distinction and proportionality, see, e.g., Andreea Manea, "Too Early To Tell? The (Un)Lawfulness of Israeli Attacks: The Case of the Jabalia Refugee Camp", *EJIL: Talk!*, 4 November 2023, available at: https://www.ejiltalk.org/too-early-to-tell-the-unlawfulness -of-israeli-attacks-the-case-of-the-jabalia-refugee-camp/; cf Luigi Daniele, "A Lethal Misconception, in Gaza and beyond: Disguising Indiscriminate Attacks as Potentially Proportionate in Discourses on the Laws of War", *EJIL: Talk!*, 7 November 2023, available at: https://www.ejiltalk.org/a-lethal-misconception-in-gaza-and-beyond-disguising-indis criminate-attacks-as-potentially-proportionate-in-discourses-on-the-laws-of-war/.
[89] See, e.g., Jens D. Ohlin, "Targeting and the Concept of Intent", *Michigan Journal of International Law*, Vol. 35, 2013, pp. 85–91.
[90] A. Haque, above note 57, p. 155; A. Zemach, above note 55, p. 74.

human rights law[91] or treated as *lex specialis*.[92] Similarly, debates over customary status often turn to which counts for more: military manuals or the actual practice of States?[93] Here, this paper echoes Ingo Venzke's caution against the infinite regress of rules justifying rules,[94] in which the answer to sources and applicable rules would ultimately come down to one's subjective values and beliefs.[95] Indeed, as section 2.4 highlights, the deeper problem with interpretative indeterminacy is that the interpretative community of LOAC/IHL is sharply divided. Rendering any alignment with LOAC/IHL as value-laden, akin to ethics or intent alignment. This poses a significant challenge to step 2 of the hypothetical scenario: whether the TFAI will consider international human rights law as a relevant rule for interpreting LOAC/IHL targeting principles would likely depend on the AI designers' *ex ante* ethical or ideological convictions.

Secondly, legal indeterminacies can arise from multiple equally valid interpretations of the same legal rule, posing specific challenges to step 2 of the hypothetical scenario. Under the law of treaties, there is no hierarchy between interpretations in good faith, according to the treaty's ordinary meaning, object and purpose, and in its context.[96] Take the prohibition on weapons causing 'superfluous injury or unnecessary suffering' as an example. NATO Member States have adopted different interpretations of the same terms in their military manuals based on different interpretative methods.[97] This has already created difficulties for NATO in agreeing on a sufficiently recognised list of prohibited weapons.[98] In light of the lack of interpretative consistency even within a highly synchronised military alliance, it is

---

[91] For a general overview of the principle of systemic integration, see, e.g., Campbell McLachlan KC, *The Principle of Systemic Integration in International Law*, Oxford University Press, Oxford, 2024; for an application of systemic integration to LOAC/IHL, see, e.g., A. Haque, above note 57, p. 160.

[92] For an elaboration of *lex specialis* vs *lex generalis*, see, e.g., Martti Koskenniemi, "Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law - Report of the Study Group of the International Law Commission", UN Doc. A/CN.4/L.682 and Add.1, paras. 88–107; On the view of LOAC/IHL as *lex specialis*, see, e.g., Michael N Schmitt, "The Interpretive Guidance on the Notion of Direct Participation in Hostilities: A Critical Analysis", *Harvard National Security Journal*, Vol. 1, 2010, p. 43.

[93] An illustrative example of the disagreement on the source and method of LOAC/IHL is the controversies regarding the ICRC's study of customary international law, D. Luban, above note 16, p. 325; see, e.g., Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, *Vol. 1: Rules*, Cambridge University Press, Cambridge, 2005 ("ICRC Customary Law Study"); and the US response, see John B. Bellinger III and William J. Haynes II, "A US Government Response to the International Committee of the Red Cross Study Customary International Humanitarian Law", *International Review of the Red Cross*, Vol. 89, 2027, p. 315.

[94] Ingo Venzke, *How Interpretation Makes International Law: On Semantic Change and Normative Twists*, 1st edn, Oxford University Press, Oxford, 2014, p. 30.

[95] Marko Milanović, *Extraterritorial Application of Human Rights Treaties: Law, Principles, and Policy*, Oxford University Press, Oxford, 2011, pp. 252–256; the same choice was also observed in the context of conflicts between trade law and patent law, see J. Rochel, above note 36, p. 22.

[96] Vienna Convention on the Law of Treaties, 1155 UNTS 331, adopted 23 May 1969 (entered into force 27 January 1980), Art. 31.

[97] A. Bellal and S. Casey-Maslen, above note 71, p. 187.

[98] See, e.g., M. Schmitt, above note 82.

foreseeable that a divergent approach to interpreting the same AI-Guiding Treaty could still lead to significant legal risk and controversy that TFAI proponents seek to avoid.

It is important to note that indeterminacy in interpreting legal norms is often intertwined with previous debates on the sources of LOAC/IHL, including the legitimacy of non-state actors to interpret treaties. Should the TFAI consider the ICRC's study on customary IHL as a subsidiary means of interpretation? While humanitarian lawyers advocate a more pluralistic approach that involves non-state actors,[99] military lawyers have carefully guarded the rule-making power for States and States only.[100] A similar ongoing debate exists on the meaning of "direct participation in hostilities" in AP I.[101] The ICRC's *Interpretative Guidance*[102] was heavily criticised by US military lawyers for overemphasising the status of military manuals and ignoring actual state practices.[103] In other words, before the TFAI could apply the distinction in the context of non-international armed conflict, AI designers are already required to take a stance *ex ante* on what counts as "direct participation in hostilities".[104] This interpretative indeterminacy was furthered by the decentralised nature of the international legal system, in which international law is produced, interpreted, implemented, and enforced by the very actors it seeks to bind. Even if AI could reliably predict how courts would decide cases,[105] the dataset of international war crimes trials is arguably too small to be representative.[106] The cases are also internally contradictory (i.e. the famous "overall control"[107] versus "effective control"[108] divide on "internationalisation" of armed conflicts) and shaped by

[99] See, e.g., the role of ICRC, judicial bodies, and highly qualified publicists, in Sandesh Sivakumaran, "Making and Shaping the Law of Armed Conflict", *Current Legal Problems*, Vol. 71, 2018, pp. 124–128.
[100] Michael Schmitt and Sean Watts, "State Opinio Juris and International Humanitarian Law Pluralism", *International Law Studies*, Vol. 91, 2015, pp. 209–210.
[101] AP I, Arts.13(3) and 51(3).
[102] ICRC and Nils Melzer, "Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law", available at: https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/icrc-002-0990.pdf.
[103] For a discussion on the 'revolving door' problem, see, e.g., Schmitt and Watts, above note 100, p. 184; William Boothby, "Direct Participation in Hostilities - A Discussion of the ICRC Interpretive Guidance", *Journal of International Humanitarian Legal Studies*, Vol. 1, 2010, p. 143; Kenneth Watkin, "Opportunity Lost: Organized Armed Groups and the ICRC "Direct Participation in Hostilities' Interpretative Guidance", *International Law and Politics*, Vol. 42, 2010, p. 641.
[104] For how "architectural choices" might harbour hidden jurisprudential assumptions about the nature of international law, see, generally, J. Rochel, above note 36, pp. 13–17.
[105] Oliver Wendell Holmes, "The Path of the Law", *Harvard Law Review*, Vol. 10, 1897, pp. 457–458; for its application to LFAI, see C. O'Keefe and others, above note 26, p. 125.
[106] One could compare this with the large number of cases and arbitral awards associated with trade and investment, and thousands of decisions from the European Court of Human Rights, see Ashley Deeks, "High-Tech International Law", *The George Washington Law Review*, Vol. 88, 2020, pp. 624–625.
[107] ICTY, *Prosecutor v Tadić, Case No. IT-94-1-A,* Judgment (Appeals Chamber), 15 July 1999, para. 131.
[108] International Court of Justice, *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Judgment, ICJ Reports 1986, para. 115.

deliberate judicial ambiguity.[109] Reliance on the "semantic authority" of international courts and tribunals is *per se* insufficient to resolve indeterminacy, particularly where fault lines are drawn not by the quality of legal arguments but by underlying values and contextual differences.[110]

Thirdly, another source of indeterminacy in LOAC/IHL arises from the existence of discretion.[111] Here, an important premise of TFAI is that a better understanding of the law improves compliance. This leaves out an important question: how TFAI should apply the law to the facts. Indeed, how we distinguish a question of law from a question of fact is already indeterminate, as it is largely based on judicial convention.[112] Moreover, research shows how one labels data on AI behaviour can itself create path dependence independent of the AI design-architecture.[113] Here, it must be noted that the compliance of some LOAC/IHL rules exclusively depends on the finding of facts, such as a specific intent to deceive the enemy for "perfidy",[114] or the determination of a prior violation of LOAC/IHL for belligerent reprisal.[115] Taken together, this means that the process of fact-finding would have already challenged TFAI's assumptions of "objective determinacy" and "neutral interpretation", which underpin the premise that clarity in law alone could secure legal compliance for agentic AI.

As stated in section 2.1 of this paper, even when a rule is linguistically determinate, it could still produce an indeterminate result when applied to a particular factual scenario. As Zurek and van Engers acknowledged, the efforts to

---

[109] See the *Gotovina et al* case discussed below. For another famous example of 'strategic ambiguity' adopted by the ICJ, see International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, ICJ Reports 1996.

[110] 'Judicial lawmaking is an inevitable byproduct in the practice of interpretation by institutions with a certain degree of semantic authority.' See, I. Venzke, above note 94, p. 136.

[111] S. Meier, above note 46, p. 50.

[112] Emad Atiq, "Legal vs. Factual Normative Questions & the True Scope of Ring", *Notre Dame Journal of Law, Ethics & Public Policy*, Vol. 32, 2018, p. 106.

[113] See, generally Aparna Balagopalan and others, "Judging Facts, Judging Norms: Training Machine Learning Models to Judge Humans Requires a Modified Approach to Labeling Data", *Science Advances*, Vol. 9, 2023, available at:https://doi.org/10.1126/sciadv.abq0701.

[114] For a striking recent illustration of the problem, see Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Articles of War*, 5 December 2025, available at: https://lieber.westpoint.edu/can-ai-teach-itself-abuse-ihl-protected-indicators/. Kwik and Wiese identified at least three novel indeterminacies in applying existing LOAC/IHL provisions to this scenario: (1) whether operator foreseeability matters, (2) what counts as an 'indicator' in cyberspace, and (3) whether it matters if the deceived party is itself an AI. See also, Dora Velenczei, "Learned Deception in AI-Enabled Lethal Autonomous Weapons Systems", *Asia-Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare*, 2026, p.17. See further Samuel White, Dora Velenczei and Jonathan Kwik, 'AI-on-AI Perfidy and the Law of Armed Conflict' *Articles of War,* 8 June 2026, available at: https://lieber.westpoint.edu/ai-on-ai-perfidy-law-armed-conflict/.

[115] It is, at least, uncontroversial that belligerent reprisal is allowed in relations to the methods and means of warfare in international armed conflicts, see Frits Kalshoven and Liesbeth Zegveld, *Constraints on the Waging of War: An Introduction to International Humanitarian Law*, 4th ed., Cambridge University Press, Cambridge, 2011, pp. 156–159.

formalise LOAC/IHL rules ultimately require the simplification of values into numbers and the reduction of subjective uncertainty into mathematical probabilities.[116] In the field of LOAC/IHL, there has been strong resistance to any quantitative standards that would constrain the discretion of military commanders. Instead, the assessment of compliance with distinction and proportionality largely depends on qualitative assessment, informed by certain unspoken conventions of human commanders.[117] The principle of distinction in AP I illustrates how, despite its relatively determinate textual formulation, the rule could still produce considerable indeterminacy in practice.[118] In the ICTY case of *Gotovina et al*, the Trial Chamber controversially held that any launch of a projectile landing more than 200 metres from a military objective would be deemed an indiscriminate attack.[119] While the Appeal Chamber overturned the 200-metre standard, it declined to articulate any alternative to define the level of accuracy for a weapon system.[120] A similar case-by-case analysis is also adopted to assess the proportionality of the attack, in which there is no hard quantitative standard for how many civilian deaths would be considered excessive.[121] Instead, commanders are told to weigh factors such as the target's military importance, the density of the civilian population, the likely incidental effects, and the timing of the attack.[122] Apart from the courts, Ashley Deeks admitted that militaries were also staunchly against the idea of having some "magic number" for proportionality,[123] and they preferred to leave the precise justification opaque behind the "based on all of the facts and circumstances" formula.[124]

While both courts and military lawyers emphasise the importance of human judgment over abstract quantifiable standards, it is important to notice that both military and humanitarian lawyers agreed that the commander's evaluation is still constrained by certain unspoken conventions of "reasonableness".[125] For example,

---

[116] Tomasz Zurek and Tom Van Engers, "The Computational Representation of Rules of International Humanitarian Law in Military Autonomous Devices: Issues, Problems, and Controversies", in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, TMC Asser Press, The Hague, 2026, pp. 245–246.

[117] There is arguably an unspoken agreement that there is no agreement. See Major General (Ret.) Walter B Huffman, "Margin of Error: Potential Pitfalls of the Ruling in the *Prosecutor v. Ante Gotovina*", *Military Law Review*, Vol. 211, No.1, 2012, p. 17, speaking of 'legal standards that are often intentionally and patently imprecise.' See also the discussion on 'strategic ambiguity' in section 2.5 of this paper.

[118] A. Bellal and S. Casey-Maslen, above n. 71, p. 122.

[119] ICTY, *Prosecutor v Gotovina et al,* Case No. IT-06-90-T, Judgment (Trial Chamber), 15 April 2011, para. 1899.

[120] ICTY, *Prosecutor v Gotovina et al,* Case No. IT-06-90-A, Judgment (Appeals Chamber), 16 November 2012, para. 83.

[121] David Kennedy, *Of War and Law*, Princeton University Press, Princeton, 2006, p.143.

[122] A. Bellal and S. Casey-Maslen, above note 71, p. 160.

[123] Israel has reportedly consulted a separate seven-member committee to determine acceptable civilian casualties per dead terrorist, and a ratio of 3.14 civilians to one dead terrorist was suggested, with adjustments if the civilians were children. Deeks, however, deemed it undesirable. See A. Deeks, above note 21, p. 58 at n. 55.

[124] *Ibid*., p. 58.

[125] See, generally, Michael Schmitt, "Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance", *Virginia Journal of International Law*, Vol. 50, No. 4, 2010, p. 795.

Michael Schmitt revealed that, during the Iraq War, an attack causing more than 30 civilian casualties would require approval from the US Secretary of State.[126] The discretion of military commanders is thus collectively and locally circumscribed by a "belt of restriction", or what Ronald Dworkin described as the "hole" in the middle of a doughnut.[127] If one accepts the Dworkinian 'doughnut' metaphor, it is still possible for an AI system to apply the principle of proportionality by adopting certain parameters circumscribing the margin of discretion. Tomasz Zurek and van Engers, for example, represent the principle of proportionality through a computational equation, which includes a proportionality coefficient that represents the view of a reasonable commander.[128] This coefficient is a blatant admission that they saw it as impossible to code or predefine the indeterminate concept of "reasonableness" in an algorithm, but they also suggested that machines could perhaps "learn" about what human commanders would consider as 'reasonable' either through data-based or knowledge-based paradigms.[129]

Both approaches have a downside. The data-based approach requires labelling, which risks generalising to cases that might not be comparable.[130] The data-based approach risks losing case-specific nuances that affect the decision, such as the quality of civilians (e.g., children and women), during generalisation.[131] The knowledge-based approach also requires a central repository of knowledge and an inference engine to derive new information via logic, rules, or constraints. Although past data on military operations could well create this repository, this could skew the "reasonableness" assessment toward a specific military culture with a predetermined view of the trade-off between "military necessity" and "human dignity".[132] Here, the risk of implementing TFAI in military targeting lies in the fact that, while TFAI may not significantly reduce the legal risks and political costs inherent in contested

[126] Michael N. Schmitt, "Precision Attack and International Humanitarian Law", *International Review of the Red Cross*, Vol. 87, 2005, pp. 457–458. However, one could question whether such a requirement was not also driven by considerations external to institutional culture, such as purely strategic interests in maintaining local support for the war effort.
[127] R. Dworkin, above note 35, p. 31.
[128] T. Zurek and T. Van Engers, above note 116, p. 249.
[129] *Ibid*., p. 258.
[130] The same observation is identified as a weakness of LLMs in case-based legal reasoning, see Ha Thanh Nguyen and others, "LLMs for Legal Reasoning: A Unified Framework and Future Perspectives", *Computer Law & Security Review*, Vol. 58, 2025, p. 7.
[131] See, e.g., Ari Spiesberger and others, "Soft Contamination Means Benchmarks Test Shallow Generalization", *arXiv*, 2026, available at: https://doi.org/10.48550/ARXIV.2602 .12413; cf Ryan Moulton, "The Many Ways That Digital Minds Can Know", *Ryan Moulton's Articles*, 28 June 2023, available at: https://moultano.wordpress.com/2023/06/28/the-many-ways-that-digital-minds-can-know/. Moulton argued that the distinction between generalisation and memorisation may be 'theoretically moot' when what matters is behavioural reliability and interpretability.
[132] Deeks, for example, suggested that this could be constructed based on the vast amount of data that the US military possessed, A. Deeks, above note 21, p. 47; see also, the existence of a particular version of IHL as a result of one state's distinctive military culture, Naz K. Modirzadeh, ""Violent, Vicious, and Fast": LSCO Lawyering and the Transformation of American IHL", Harvard National Security Journal, Vol. 17, 2025; J. Rochel, above note 36, p. 19.

military decision-making, it removes the human burden of justifying and bearing responsibility for a particular legal interpretation.[133]

A further challenge concerns risk-taking by military commanders.[134] In step 4 of the hypothetical scenario, TFAI must include a margin of appreciation for acceptable risks of LOAC/IHL violations.[135] Here, the question of "rigour" of TFAI becomes relevant.[136] Maas and Olasunkanmi suggested that the TFAI proposal requires a certain acceptance of the concept of "efficient breach"[137] so that the AI system can undertake a certain degree of legal risk.[138] Yet "efficient breach" of LOAC/IHL is arguably the hardest to justify amongst all branches of international law. First, this will contradict one of TFAI's main objectives: to align AI behaviour with peremptory norms of international law.[139] This is especially relevant to the principles of distinction and proportionality, which, ironically, are widely recognised as enjoying peremptory status notwithstanding their indeterminacies. Second, the irreducibility of military decisions to cost-benefit analysis undermines any assessment of whether the breach would be considered as "efficient".[140] It will also upset reciprocity between belligerents,[141] which could lead to an escalation of belligerent reprisals, potentially undermining the entire structure of LOAC/IHL.[142] Instead of admitting any "breach", States and armed groups have found it more desirable to rely on existing justification based on a broad reading of military advantage and/or necessity. In other words, LOAC/IHL is arguably one of the legal fields that demands the highest degree of compliance. Yet this demand is not satisfied by a mechanical calculation of whether a breach would be efficient. Rather, it is met through a reasonableness assessment of the risk-taking discretion exercised by peer military commanders, which remains difficult for AI systems to fully replicate.

In sum, fact-assessment indeterminacy undermines TFAI's "neutral interpretation" thesis as the application of many LOAC/IHL's fundamental

---

[133] See section 2.4 of this paper.
[134] Zurek and van Engers argued that this would involve the development of a sound calculus for probabilistic reasoning that could account for the interdependence of root facts, see T. Zurek and T. Van Engers, above note 116, pp. 261–262.
[135] Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies*, TMC Asser Press, The Hague, 2024, p. 70.
[136] M. Maas and T. Olasunkanmi, above note 3, p. 33.
[137] Eric A. Posner and Alan O. Sykes, "Efficient Breach of International Law: Optimal Remedies, 'Legalized Noncompliance,' and Related Issues", *Michigan Law Review*, Vol. 110, 2010, p. 243; cited in M. Maas and T. Olasunkanmi, above note 3, p. 33 at n. 134.
[138] It is argued that the scope of acceptability could be defined by a treaty, through arbitral body adjudication, or in other texts, see M. Maas and T. Olasunkanmi, above note 3, p.32.
[139] Draft Conclusion 2 in International Law Commission, above note 23.
[140] The limit of 'efficient breach' theory in *jus in bello* was recognised, see Posner and Sykes, above note 137, p. 283.
[141] Michael N. Schmitt, "Asymmetrical Warfare and International Humanitarian Law", *Air Force Law Review*, Vol. 62, 2008, pp. 41–42.
[142] For the scope of "belligerent reprisals", see F. Kalshoven and L. Zegveld, above note 114, pp. 156–159.

principles ultimately involves difficult ethical judgment, such as the balancing of a combatant's life with that of the enemy's civilian. Moreover, its reliance on "efficient breach" to accommodate discretion further conflicts with the doctrinal status of certain indeterminate LOAC/IHL rules. The next section shows that these difficulties have a deeper structural origin, linking them to the conflicting role of LOAC/IHL in simultaneously legitimising and restraining States' use of violence. Further, as section 2.5 will argue, the legal indeterminacies surveyed above, such as the absence of hierarchy among interpretations, the contested status of customary rules, and the unresolved boundary between distinction and proportionality, are precisely those questions that States and tribunals have repeatedly declined to settle. This suggests that legal indeterminacy is not a residual gap awaiting clarification but an actively maintained feature of the LOAC/IHL.

3. Structural indeterminacy

Structural indeterminists hold that international law exhibits a specific kind of indeterminacy because, compared to domestic law, it lacks a coherent account of its own legal validity.[143] David Kennedy and Martti Koskenniemi argued that international law is structurally indeterminate because its authority rests on both sovereign consent (i.e. legal positivism) and non-consensual norms (i.e. natural law). Both bases are logically contradictory and mutually exclusive.[144] Yet, structural indeterminists observed that a strong argument in international law must oscillate between consent-based discourse (apology) and non-consensual discourse (utopia) to avoid revealing international law's ultimate contradiction.[145] This created what Koskenniemi called a perceived determinacy of international law through the "culture of formalism".[146] The structural indeterminacy school poses a profound challenge to the second TFAI assumption (i.e. the "objective determinacy thesis"). In this sense, the indeterminacy of international law is not a technical challenge but a conceptual collision with the fundamental idea of aligning AI with LOAC/IHL treaties as if the latter represent an 'appropriate, concrete, and certified' standard.[147]

---

[143] S. Meier, above note 46, pp. 51–54.

[144] See, generally, David Kennedy, *International Legal Structures*, Baden-Baden, Nomos-Verl-GmbH, 1987, pp. 29–54; David Kennedy, "Theses about International Law Discourse", *German Yearbook of International Law*, Vol. 23, 1980, p. 374.

[145] David Kennedy, "A New Stream of International Law Scholarship", *Wisconsin International Law Journal*, Vol. 7, 1988, p. 38; D. Kennedy, above note 121, p. 107; see also Martti Koskenniemi, *From Apology to Utopia: The Structure of International Legal Argument*, Cambridge University Press, Cambridge, 2005.

[146] See, generally Martti Koskenniemi, *The Gentle Civilizer of Nations the Rise and Fall of Modern International Law 1870-1960*, Cambridge University Press, Cambridge, 2009.

[147] For a similar point that indeterminacy is an inherent feature of the law, see, e.g., J. Rochel, above note 36, p. 25.

The structural indeterminists highlighted the dual role of LOAC/IHL in both legitimising and restricting the use of State violence.[148] As such, this structural contradiction is central to the identity of LOAC/IHL: Frédéric Mégret, for instance, advanced Koskenniemi's framework [149] by arguing that the indeterminacy of combatant classification goes to the very function of LOAC/IHL,[150] whose legal meaning historically hinged on unstated social-ideal types, such as "chivalry", the standard of "civilisation", and the concept of "restrained masculinity".[151] David Kennedy argued that the field's structural indeterminacy has spawned two fundamentally distinctive interpretative communities: military lawyers and humanitarian lawyers, with the former prioritising military necessity and the latter emphasising human dignity.[152] Kennedy's view on LOAC/IHL indeterminacy was nuanced, as he accepted that, even in the absence of a determinable standard, there were easy and hard cases that even a military commander could instinctively identify.[153] Instead of some objective internal rules, Kennedy believed that discretion was only circumscribed by one's sense of professional responsibility.[154]

Here, the most important distinction between structural indeterminists and others is in their focus on the significance of the indeterminacy rather than its degree.[155] For instance, Kennedy highlighted that LOAC/IHL mattered the most not in "easy" and "hard" cases, but "in the great run of situations that fall in between."[156] As such, because indeterminacy in LOAC/IHL generated extensive grey areas in the day-to-day targeting decisions, the stable legal core that positivists rigorously defend became peripheral in practice. Take, for example, the assessment of distinction and proportionality in steps 2 and 3 of the hypothetical scenario: structural indeterminists will argue that any search for determinacy will fail because a great amount of assessment will oscillate between the two poles of "apology" (i.e. the LOAC vision) and "utopia" (i.e. the IHL vision).

---

[148] On the dual role of international law in legitimising violence, see Ian Hurd, *How to Do Things with International Law*, Princeton University Press, Princeton, 2017, pp. 82–102.

[149] Frédéric Mégret, "The Apology of Utopia: Some Thoughts on Koskenniemian Themes, with Paritcular Emphasis on Massively Institutionalized International Human Rights Law", *Temple International & Comparative Law Journal*, Vol. 27, 2013, pp. 455, 489.

[150] See, generally, Frédéric Mégret, "From "Savages" to "Unlawful Combatants": A Postcolonial Look at International Humanitarian Law's "Other"", in Anne Orford (ed.), *International Law and its Others*, 1st ed., Cambridge University Press, Cambridge, 2006, pp. 265–317.

[151] See, e.g., generally Frédéric Mégret, "The Laws of WAR and the Structure of Masculine Power", *Melbourne Journal of International Law*, Vol. 19, 2018, p. 200; Frédéric Mégret, "Another Look at the Gendered Constitution of the Laws of War", *Journal of International Humanitarian Legal Studies*, Vol. 14, 2023, p. 304.

[152] See generally, D. Kennedy, above note 121, pp. 28, 37–39, 85–88, 129–130.

[153] *Ibid*., pp. 143–144.

[154] *Ibid*., p. 145.

[155] In this regard, this paper departs from Meier's framework, which primarily discusses whether the law is globally or locally indeterminate; see, cf., S. Meier, above note 46, p. 59.

[156] D. Kennedy, above note 121, 144.

On the surface, structural indeterminists reveal the inherent argumentative structure of international law, which LLM models can easily "imitate" without understanding its broader societal function. Here, the fundamental incompatibility between TFAI and LOAC/IHL lies in AI's inability to understand and distinguish between the two irreconcilable purposes of LOAC/IHL.[157] Recent research has shown that LLMs can only approximate, rather than fully capture, the complexity of policy-based and principled legal reasoning.[158] In this sense, AI cannot understand that legal interpretation serves a purpose beyond finding the meaning of the norm. Namely, creating a "culture of formalism" that persuades decision-makers to distance themselves from their preferences and consider the generality of their positions.[159] Even if AI developed sufficient moral reasoning capacity to engage in such interpretation,[160] the ultimate ethical reference points of LOAC/IHL would themselves remain heavily contested between military and humanitarian lawyers, a point developed in section 2.4 of this paper.

Even if one assumes that LLM models can imitate 100% of how a capable LOAC/IHL advocate argued a case by oscillating between "apology" and "utopia" in steps 2 and 3 of the hypothetical scenario. TFAI still cannot confer the same legitimacy in 'arguing' as human lawyers in step 5 of the scenario. By requiring the AI to follow key LOAC/IHL treaties automatically, the TFAI will act upon its own interpretation without this epistemic distance, and without the awareness that a publicly defensible position cannot go all the way toward either "apology" or "utopia". TFAI thus faces a categorical mismatch with LOAC/IHL: it imitates legal argumentation without performing it through discursive practice. It is perhaps ironic that, at the age of AI, the major contribution of structural indeterminacy theory is precisely to highlight the irreplaceable role of human international lawyers.[161] Yet the structural account, while powerful, leaves an important question open: If LOAC/IHL is structurally indeterminate because it must simultaneously serve the irreconcilable purposes of legitimising and restricting State violence, why have States not, over the long arc of the twentieth and twenty-first centuries, resolved the contradiction in favour of one purpose or the other? The answer, as section 2.5 will argue, is that the structural contradiction itself is what States have agreed to preserve.

---

[157] Nasir Qadri, "Is Artificial Intelligence Quietly Killing International Law", *The Contrapuntal*, 30 December 2025, available at: https://thecontrapuntal.com/is-artificial-intelligence-quietly-killing-international-law/.
[158] H. Nguyen *et al.*, above note 130, p. 10.
[159] M. Koskenniemi, above note 146, p.501.
[160] M. Chinen, above note 9, p.379.
[161] Frédéric Mégret shows how professionals manage structural indeterminacy through institutional norms, professional socialization, and disciplinary gatekeeping, see generally, Frédéric Mégret, "Thinking about What International Humanitarian Lawyers "Do: An Examination of the Laws of War as a Field of Professional Practice", in Wouter Werner, Marieke de Hoon and Alexis Galán (eds), *The Law of International Lawyers*, Cambridge University Press, Cambridge, 2017, pp. 265-296.

4. Practice (in)determinacy

Practice (in)determinists argue that the structural indeterminists overstate the indeterminacy in the practice of international law, because any structural indeterminacy can be significantly mitigated by the common practice and beliefs of the "interpretative community".[162] This paper terms this school of thought the "practice (in)determinacy" school.[163] Andrea Bianchi argued that legal meaning is not open-ended but constrained by the interpretive community's shared "background conventions" about how rules should be interpreted.[164] His view was echoed by Jean d'Aspremont's account of international law as a "self-perpetuating belief system"[165] and by Severin Meier, who believes that the key question of determinacy lies in what is "accepted" by the interpretive community as valid at a particular time.[166] In the eyes of practice (in)determinists, fundamental categories, such as combatants and civilians in the context of non-international armed conflicts, are not indeterminate because of the incompleteness of the rules,[167] but because of the contrasting "background conventions" shared by two interpretative communities (i.e. whether LOAC/IHL should uphold military necessity or humanitarian considerations).[168] This explains why, contrary to the structural indeterminists' arguments that international law is 'globally indeterminate', there are certain areas of LOAC/IHL that are more determinate than others, because they do not touch on the fundamental conflicts between military and humanitarian lawyers over what LOAC/IHL ought to be.

In the context of LOAC/IHL, the key issue is whether the divide among interpretative communities can be bridged. David Kennedy, for instance, saw the ambiguity of LOAC/IHL as an achievement, as it provides "a common vocabulary—of rules and exceptions, broad standards and disputed interpretation—for military and humanitarian professionals assessing the legitimacy of warfare."[169] Kennedy also warned of the danger of political competition between two

---

[162] See, e.g., famously, Oscar Schachter, "Invisible College of International Lawyers", *Northwestern University Law Review*, Vol. 72, 1977, p. 217.
[163] This school of thought was partly inspired by Stanley Eugene Fish, *Is There a Text in This Class ? The Authority of Interpretive Communities*, Harvard University Press, Cambridge MA, 1980; and partly by the constructivist school in international relations, see Peter M. Haas, "Introduction: Epistemic Communities and International Policy Coordination", *International Organization*, Vol 46, No. 1, 1992, pp. 1-35.
[164] See, generally, Andrea Bianchi, "Textual Interpretation and (International) Law Reading: The Myth of (in)Determinacy and the Genealogy of Meaning", in Pieter HF Bekker, Rudolf Dolzer and Michael Waibel (eds)., *Making Transnational Law Work in the Global Economy*, Cambridge University Press, Cambridge, 2010.
[165] See, generally, Jean d'Aspremont, *International Law as a Belief System*, Cambridge University Press, Cambridge, 2017.
[166] S. Meier, above note 46, p. 86, arguing 'military lawyers' as constituting one distinctive interpretive community.
[167] A. Haque, above note 57, pp. 135–138.
[168] S. Meier, above note 46, p.151.
[169] D. Kennedy, above note 121, p. 42.

interpretative communities over the "ownership" of interpretation, which would inevitably erode trust and undermine the legitimacy of the LOAC/IHL enterprise, even in the eyes of practitioners.[170] A similar view was shared by Sandesh Sivakumaran, who highlighted how the "silence" of States has elevated the ICRC, courts, and academics as the day-to-day law-makers,[171] an insight echoed even by military lawyers.[172] While David Luban shared Kennedy's view on the importance of a common vocabulary,[173] he observed that over the past decade, the divide extended beyond mere disagreements over specific LOAC/IHL norms to encompass sources and methods, and that "the modes of argument themselves differ systematically, because the LOAC and IHL visions represent coherent systems of jurisprudence that turn on different 'Archimedean points': military necessity or human dignity."[174] To summarise, the practice determines the identity of LOAC/IHL not by its value as an objective benchmark but by its role as a common discourse that performs a 'world-making' function, bridging parties with irreconcilable worldviews.[175]

The first concern that the practice (in)determinist would raise about the TFAI proposal is interpretative inclusivity.[176] A major concern could be that a smaller group of lawyers, or even AI engineers from big tech companies, could effectively "own" the discursive space while obscuring the inherently political nature of LOAC/IHL interpretations.[177] Ashley Deeks, for instance, suggested that the US should involve the ICRC in the coding of LOAC/IHL.[178] The feasibility of this proposal is questionable, as it would necessarily reveal the technical capabilities of one's agentic system during the global AI arms race. Furthermore, the question of whether AI systems can participate meaningfully in legal discursive practice depends on implementation details. If TFAI systems operate as black boxes with opaque decision-making, they clearly cannot engage in the *ex post facto* justificatory discourse that LOAC/IHL's interpretative communities require. However, if chain-of-thought reasoning is transparent and made available for public inspection and review by

[170] *Ibid.*, p.135.
[171] S. Sivakumaran, above note 99, pp. 159–160.
[172] See generally, M. Schmitt and S. Watts, above note 100, pp. 179–198.
[173] D. Luban, above note 16, pp. 317–318.
[174] Ibid p.336.
[175] For the view of international law as 'language', see, e.g., Megan Donaldson, "International Law, Language and Worldmaking", *EJIL: Talk!*, 3 December 2025, available at: https://www.ejiltalk.org/international-law-language-and-worldmaking/.
[176] This relates to the question of 'representativity' in interpretation, see J. Rochel, above note 36, p. 21.
[177] The prospects of 'perceived' automatic compliance might further marginalise the role of military lawyers in the context of military operations. See, e.g., Jack Goldsmith, "The Decaying Legal Culture in the Defense Department", *Executive Functions*, 3 January 2026, available at: https://www.execfunctions.org/p/the-decaying-legal-culture-in-the.
[178] A. Deeks, above note 21, p. 57.

different interpretative communities, some form of *ex post* discursive participation is still possible.[179]

A second, and more fundamental concern that practice (in)determinists could raise is how to reconcile TFAI's positivist understanding of legal interpretation with the discursive function of LOAC/IHL. Even with full transparency in algorithmic reasoning, there is also a practical question of whether algorithmically generated reasoning could replace discursive participation of State and military lawyers.[180] This is especially in light of the fact that the TFAI proposal sees legal interpretation through a self-contained circulatory loop. By requiring an AI system to make its own interpretation and then use it as a benchmark to restrain its very own behaviour, it replaces the existing systems of *ex ante* deliberations between human lawyers within a military or across government departments. As steps 2 to 5 of the hypothetical scenario show, what would ordinarily be the task of State and military lawyers, i.e. advising military commanders on the legality of military targeting, is becoming fully automated. Here, TFAI risks foreclosing the discursive space that makes LOAC/IHL acceptable across diverse interpretative communities. In this sense, the TFAI could create a false "performative determinacy", in which legal interpretations appear legally settled merely because other interpretative communities cannot adequately engage with them.

Finally, even with full inclusivity, AI could still exhibit default biases toward one interpretation over another due to inherent differences in cognitive and algorithmic reasoning. For example, existing AI models have been shown to favour the prosecution of crimes even with minimal evidence or in cases involving clear constitutional violations, due to the inability of AI to give the accused the benefit of the doubt or to understand the value of constitutionalism.[181] In fact, Maas and Olasunkanmi acknowledged the mismatch between human and AI judgments as a potential shortcoming of the TFAI proposal.[182] It is likely that similar unconscious bias in algorithmic reasoning could result from AI's inability to detect cultural

---

[179] Maas and Olasunkanmi suggested, for instance, creating cached databases containing all the chain-of-thought reasoning logs, which could be made publicly available, see M. Maas and T. Olasunkanmi, above note 3, pp. 32–33.

[180] The discursive aspect of interpretation is arguably an "inalienable obligation" of human lawyers, which cannot be delegated to the AI. See the metaphor of worshipping a deity in S. Soltanzadeh, above note 65, pp. 40–41.

[181] See, e.g., Rory Pulvino, Dan Sutton and JJ Naddeo, "Hiding in Plain Sight: An Empirical Study of Prosecutorial Bias in AI Legal Analysis", *Science and Technology Law Review*, Vol. 27, 2026, available at: https://doi.org/10.52214/stlr.v27i1.14543.

[182] M. Maas and T. Olasunkanmi, above note 3, p. 38; In their paper, they cited a recent experiment in which the LLM delivered judgments strongly influenced by legal precedent rather than by sympathy toward the defendant in a simulated war-crime appeal trial. See Eric Posner and Shivam Saran, "Judge AI: Assessing Large Language Models in Judicial Decision-Making", *Coase-Sandor Working Paper Series in Law and Economics*, No. 3, 2025, available at: https://chicagounbound.uchicago.edu/law_and_economics /1044.

nuances.[183] Yet, compared with State and military lawyers, who must bear moral, political, and legal responsibilities for their legal interpretations,[184] a non-human agent cannot bear responsibility at all.[185] This is especially so when LOAC/IHL, in much of its operation, functions less as an *ex ante* constraint on military action than as an *ex post* framework deployed by belligerents to defend operational choices, by humanitarian actors to contest them, by courts and tribunals to allocate responsibility, and by States to negotiate the meaning of past conduct in the conduct of future relations. To force *ex ante* resolution through TFAI is therefore not simply to constrain actions of AI. It risks eliminating the discursive space within which LOAC/IHL operates and foreclosing the future contestation through which the law's meaning is continually remade.

The four indeterminacy theories surveyed thus far, regardless of whether they are linguistic, legal, structural, or practice-based, all share a common explanatory limitation. Each describes a form that indeterminacy in LOAC/IHL takes, but none fully explains why this indeterminacy is so resistant to resolution.[186] State practice over a century has had ample opportunity to clarify linguistic ambiguities, settle interpretive disputes, resolve structural contradictions, and enforce agreements among interpretive communities. The next section argues that this is because the indeterminacies are not residual but constitutive: they are actively preserved by States as a matter of strategic choice. On this basis, this paper argues that "strategic ambiguity" serves as a "meta-indeterminacy" through which the others persist, which poses the most fundamental challenges to TFAI, which aims to anchor the legitimacy of *ex ante* legal constraints on the State's consent to a treaty.

5. Strategic ambiguity

In the context of LOAC/IHL, the persistence of indeterminacy in LOAC/IHL is not an accident of legal language, a failure of judicial coordination, or a consequence of contested interpretive communities. It is arguably a deliberate diplomatic achievement that States have actively preserved because the indeterminacy itself

---

[183] See, e.g., Dominick Kubica and others, "Can AI Read Between The Lines? Benchmarking LLMs On Financial Nuance", *arXiv*, 22 May 2025, available at: https://doi.org/10.48550/ arXiv.2505.16090.

[184] Schmitt and Watts highlighted that, even when the US government tried to remain silent on the legality of 'enhance interrogation techniques' and the detention of civilians during the War on Terror, public and political debates swirled, forcing legal debates within and between several US executive branch agencies on the interpretation of LOAC/IHL norms to come up with publicly defensible positions, see M. Schmitt and S. Watts, above note 100, pp. 203–204.

[185] An important and innovative aspect of the TFAI proposal, as compared to the LFAI, is that it does not require AI to have any kind of legal personhood, see M. Maas and T. Olasunkanmi, above note 3, pp. 76–80.

[186] Severin Meier hinted that some interpretations were accepted because they were acts of power, see, e.g., S. Meier, above note 46, pp. 145–146, citing Carl Schmitt, *Positionen und Begriffe, im Kampf mit Weimar – Genf – Versailles 1923–1939*, 4th ed., Duncker und Humblot, Berlin, 2014, p. 87; for a similar idea, see Ryan Martinez Mitchell, "Sovereignty and Normative Conflict: International Legal Realism as a Theory of Uncertainty", *Harvard International Law Journal*, Vol. 58, No. 2, 2017, p. 456.

serves their continuing interests. This paper advances "strategic ambiguity" as an original concept distinct from the more familiar concept of "constructive ambiguity" developed by Michael Byers, which referred to the use of ambiguous terms to facilitate agreement.[187] "Strategic ambiguity", as developed by this paper, broadly describes an agreement of States not to reach an agreement. This covers an agreement not to agree on the linguistic meaning of certain terms of the agreement, a complete exclusion of certain sensitive issues from the agreement, or no agreement at all.[188] As such, "strategic ambiguity" covers "constructive ambiguity" but also goes beyond it. The existence of "strategic ambiguity" in core LOAC/IHL principles has helped preserve States' continuing interests in interpretive flexibility, plausible deniability, and operational latitude.

This ambiguity effectively falsifies the TFAI's first assumption, because the production of a commonly agreed text does not necessarily reflect the genuine meeting of minds between States (i.e. a thick consent), but rather what they found acceptable in the circumstances (i.e. a thin consent).[189] In other words, it is argued that key LOAC/IHL treaties were, from their historical origins, a matter of disagreement. Since the content of LOAC/IHL norms would necessarily remain indeterminate until interpreted and applied, States could, *eadem ratione*, render certain norms indeterminate by withholding their rights to interpret or apply them.[190] The difficulty in aligning AI with "strategically ambiguous" text is precisely that TFAI would alter the structure of the agreement that States were prepared to sign by closing interpretative gaps that States themselves deliberately left open.

As international legal theorist Philip Allott famously stated: "A treaty is a disagreement reduced to writing".[191] History shows that some indeterminacies in

---

[187] The adoption of ambiguous terms to facilitate agreements corresponds to what Michael Byers referred to as "constructive ambiguity", see, e.g., Michael Byers, "Still Agreeing to Disagree: International Security and Constructive Ambiguity", *Journal on the Use of Force and International Law*, Vol. 8, No. 1, 2021, pp. 91–114; Michael Byers, "Agreeing to Disagree: Security Council Resolution 1441 and Intentional Ambiguity", *Global Governance*, Vol. 10, 2004, pp. 165–186.

[188] See, e.g., see, e.g., Anthony D'Amato, "Purposeful Ambiguity as International Legal Strategy: The Two China Problem", in Jerzy Makarczyk (ed.), *Theory of international law at the threshold of the 21st century: Essays in honour of Krzysztof Skubiszewski*, Brill Nilhoff, Leiden, 1996, pp. 109-121; but it was also applied in the WTO context, see also B.S. Chimni, "China, India and the WTO Dispute Settlement System: Towards an Interpretative Strategy", in Muthucumaraswamy Sornarajah and Jiangyu Wang (eds), *China, India and the International Economic Order*, 1st ed., Cambridge University Press, Cambridge, 2010, p. 229.

[189] The distinction between "thin" and "thick" consent can be compared to the metaphors of "thick" and "thin" state consensus. "Thin" consensus in international law refers to traditional, formal state agreement, whereas "thick" consensus involves substantive participation in informal lawmaking to ensure legitimacy. See, e.g., J. Pauwelyn, R.A. Wessel and J. Wouters, above note 30; relatedly, see also Hans J. Morgenthau, *Politics Among Nations: The Struggle for Power and Peace*, 4th edn, Alfred A Knopf, New York, 1967, p. 269.

[190] Ambiguity is arguably the "normal" condition unless States are willing to go to war because of such disagreement. See, e.g., R. Mitchell, above note 186, 470.

[191] Philip Allott, "The Concept of International Law", *European Journal of International Law*, Vol. 10, 1999, p. 43.

LOAC/IHL were created and sustained by "constructive ambiguity". A widely cited example in this regard is the Martens Clause, introduced by Russia at the 1898 Hague Peace Conference to accommodate different positions on the status of *franc-tireurs*.[192] Yet recent historical research has shown that the deliberate choice to adopt "constructive ambiguity" extended beyond the Martens Clause and became the defining feature in the discussion of irregular armed forces.[193] The open definition of "armed conflicts not of an international character" in Common Article 3, for instance, was a deliberate gap left to shield European States from legal scrutiny of their colonial counterinsurgencies.[194] Ironically, the same indeterminacy in Common Article 3 also became the basis for the Soviet Bloc and Third World States to extend POW status to fighters participating in the "wars of national liberation" under Article 1(4) of the First Additional Protocol.[195] In response, the US-led Western Bloc deliberately left the terms "engagement" and "deployment" opaque in the "carry arms openly" requirement of Article 44(3) of the same treaty,[196] allowing both permissive and restrictive readings in the granting of POW status to guerrilla fighters.[197] These "constructive ambiguities" have real-life implications. Returning to the hypothetical scenario, the evaluative terms render uncertain whether the agentic AI system could recognise the combatant status of enemy fighters who wore civilian clothes and engaged in hostilities in an armed conflict that one side considers to be a "war of national liberation".[198]

Apart from "constructive ambiguity", certain indeterminacies of key LOAC/IHL rules stemmed from the deliberate absence of any agreement, and the resulting deferral to more general principles involving human evaluative judgments. Jochnick and Normand showed that the indeterminacy in relation to the prohibition on indiscriminate attack could be traced back to the Hague laws of 1899, in which

---

[192] See, e.g., Rupert Ticehurst, "The Martens Clause and the Laws of Armed Conflict", *International Review of the Red Cross*, Vol. 37, 1997, p. 125; Antonio Cassese, "The Martens Clause: Half a Loaf or Simply Pie in the Sky", *European Journal of International Law*, Vol. 11, 2000, p. 187.

[193] *Franc-tireurs*, and *levee en masse* were however given full protection under GC III, Art 4.

[194] See, the joint efforts of the UK and France in securing the approval of the French-tailored text, which eventually became Common Article 3, see Giovanni Mantilla, *Lawmaking under Pressure: International Humanitarian Law and Internal Armed Conflict*, Cornell University Press, Ithaca, 2020, pp. 89–95; see what Van Dijk referred to as 'studied vagueness', B. van Dijk, above note 78, pp. 115–117, 122, 176–177.

[195] B. van Dijk, above note 78, pp. 322–323; The article prevented South Africa, Israel, and the US from ratifying AP I, see François Bugnion, "Adoption of the Additional Protocols of 8 June 1977: A Milestone in the Development of International Humanitarian Law", *International Review of the Red Cross*, Vol. 99, 2017, pp. 785, 792–793.

[196] G. Mantilla, above note 194, p. 132; Although the strategic ambiguity was defended by the American delegation, it became one of the main reasons why the US government ultimately declined to sign or ratify AP I. Giovanni Mantilla, "The Origins and Evolution of the 1949 Geneva Conventions and the 1977 Additional Protocols", in Matthew Evangelista and Nina Tannenwald (eds.), *Do the Geneva Conventions Matter?*, Vol. 1, Oxford University Press, Oxford, 2017, pp. 59–60.

[197] In this regard, the US provided the ICRC with its template used in Vietnam (and French in Algeria), which subsequently became the basis of the draft, see, e.g., G. Mantilla, above note 194, p. 157.

[198] See, e.g., the debates regarding the "continuous combat function" and the thresholds for "direct participation of hostilities" proposed by the ICRC, and the push-back by States and military lawyers, A. Bellal and S. Casey-Maslen, above note 71, pp. 138–139, 147–149.

the meaning of "undefended" in the prohibition of bombardment was deliberately undefined.[199] During the drafting of the Geneva Conventions, the legality of aerial bombing and nuclear weapons was heavily discussed,[200] with the UK and the US emphasising military necessity, and the Soviet Union invoking humanitarianism.[201] The omission of the principle of distinction in the Fourth Geneva Convention was thus not accidental, but the very result of compromises between the Great Powers and the ICRC.[202] At the Diplomatic Conferences from 1974 to 1977, there was a similar consensus against debating the legality of nuclear weapons. Even when the ambiguity surrounding "aerial bombardment" was significantly narrowed,[203] not least because of Vietnam's involvement in drafting after 1975,[204] several "constructive ambiguities" persist in the principle of proportionality (i.e. what counts as "excessive" or "proportionate"), leading to both a restrictive and permissive reading.[205]

Here, one could connect the reluctance of States and international courts to lay out clearly defined standards to the need to preserve "strategic ambiguity", which on one hand, formed the basis of consent by some Nuclear States to accede to key LOAC/IHL treaties; and on the other hand, for non-nuclear States and humanitarian organisations to argue against the use of nuclear weapons based on general LOAC/IHL principles. In this regard, "strategic ambiguity" ensures that LOAC/IHL can secure "some" compliance from a larger group of States than full legal determinacy among a much smaller group of States.

Arguably, "strategic ambiguity" in LOAC/IHL presents challenges to the TFAI proposal in several ways. First, "strategic ambiguity" undermined TFAI's use of State consent as a legitimacy anchor. Ryan M. Mitchell, for example, argued that one of the important attributes of "sovereignty" is the authority to put an end to epistemic indeterminacy by instituting binding norms and judgments.[206] TFAI's

---

[199] Chris A. Jochnick and Roger Normand, "The Legitimation of Violence: A Critical History of the Laws of War", in Michael Schmidt (ed.), *The Development and Principles of International Humanitarian Law*, London, Routledge, 2017, pp. 80–81; The Great Powers also failed to agree on the legality of submarine warfare, and naval bombardment, see B. van Dijk, above note 78, pp. 203–204.
[200] See, generally B. van Dijk, above note 78, pp. 197–251.
[201] The UK and the US were successful in blocking the matter of air-nuclear warfare from the final drafts by blocking the Soviet proposal on various procedural grounds, see *ibid*., pp. 226–248.
[202] The final draft provided few means to stop or limit air-nuclear warfare, except for prohibiting human shields, which served the US interests in protecting fallen pilots, *ibid*., pp. 249–250.
[203] Giovanni Mantilla, "Stealth Change: How the United States and the United Kingdom Embraced International Humanitarian Limits on Nuclear Use", *Security Studies*, Vol. 1, 2026, p. 18; Proportionality was first proposed in the context of the 1929 revision of the Hague Rules to address aerial bombings, B. van Dijk, above note 78, p. 209; The UK-US objection led to the adoption of the non-binding New Delhi Rules in 1957, B. van Dijk, above note 78, pp. 324–325.
[204] Tim McCormack, "Negotiating the Two Additional Protocols of 1977: Interview with the Right Honourable Sir Kenneth Keith", in Suzannah Linton, Tim McCormack and Sandesh Sivakumaran (eds.), *Asia-Pacific Perspectives on International Humanitarian Law*, 1st ed., Cambridge University Press, Cambridge, 2019, p. 25.
[205] See *Nuclear Weapons*, above note 109, paras. 55, 95.
[206] R. Mitchell, above note 186, p. 423.

deepest problem is that it usurps the sovereign role without admitting it. It also allows a smaller group of lawyers and computer scientists to deliberately close interpretative gaps left open by States, without the legitimacy structure that normally accompanies such a function.

Secondly, what "strategic ambiguity" shows is that "thin consent" at the textual level is incapable of producing "thick consent" at the application level. Here, the danger of relying on the State's 'thin consent' as a benchmark could lead only to performative compliance that undermines the legitimacy TFAI has gained. Scott Sullivan gave the example that if States were to clarify the standard for when it became "necessary" to detain civilians in armed conflicts, they might find it easier and more convenient to identify them as military targets. Citing Israeli military operations in Gaza as an example, Scott Sullivan warned that applying LFAI to a military context might end up "creating a legally compliant system that is causing far more damage and destruction than what we thought, as legal architects, we were constructing".[207] Similarly, Delgado highlighted the danger that an AI agent may conduct "ethically dubious but legal behaviours" that undermine public trust in the overall project of legal alignment.[208] These examples showed that, were States to specify, *ex ante*, a certain evaluative threshold, the result would be politically intolerable and would undermine the legitimacy gains that TFAI seeks to achieve. In other words, States that today comply imperfectly with a flexible standard would tomorrow be forced either to comply rigidly with a concrete standard, or to repudiate the normative object and purpose of the standard itself.

Taking their points further, one might question whether a mechanical striving for legal certainty is preferable to the kind of "world making" which LOAC/IHL has been doing. This could be seen from the successful example of ideologically opposed and geographically diverse States in agreeing to a common legal language on warfare in 1949 and 1977 respectively, ambiguity was the key to keep ideologically diverse States on board. In other words, "ambiguity" is the only way in which contested international norms can become universal, absent a global sovereign to enforce an agreement across States.[209] Instead of trading ambiguity for clarity, TFAI risks sacrificing a working legal order for a coercive technical one. As stated, TFAI might drive AI-leading powers, such as the US and China, to unilaterally encode their normative preferences regarding LOAC/IHL into their own AI architectures and present their interpretations of LOAC/IHL as *fait accompli* before their peer-adversaries close the interpretative gap.[210] TFAI thus forces AI-leading powers into a prisoner's dilemma: adopt the permissive military interpretation (and absorb political and legal challenges), embrace the restrictive humanitarian interpretation (which militaries may refuse), or attempt a compromise that satisfies neither side. Thus, from the perspective of States lagging in AI

[207] See Scott Sullivan in Institute for Law & AI, above note 24.
[208] K. Delgado, above note 41, pp. 14–15.
[209] R. Mitchell, above note 186, pp. 451–452.
[210] On interpretative diversity between the US and China, see J. Rochel, above note 36, p. 31.

development and in their capacity to fine-tune their acquired military AI systems to their own interpretations, TFAI might transform technical hegemony into legal hegemony.[211]

Therefore, "strategic ambiguity" has arguably functioned as the meta-indeterminacy that runs through and explains the other four schools. Linguistic indeterminacies persist because States prefer to keep the meaning of some definitions open-ended. Legal indeterminacies persist because no tribunal has had sufficient interest to tell States (and by extension, military commanders) how and which LOAC/IHL rules should be applied to a concrete situation. Structural indeterminacies persist because the contradictory purposes of LOAC/IHL are precisely what States failed to agree. And the divide between interpretive communities persists because it provides each State the choice of which interpretive tradition to invoke when convenient. Hence, "strategic ambiguity" is what underlies the other four indeterminacies. This presents perhaps the deepest challenge to TFAI's 'genuine agreement' thesis. States consented not to a determinate rule that future practice would clarify, but to a "superposition" that future practice was meant to preserve. To resolve such ambiguity, therefore, is not to anchor the restraints on the actions of AI systems in States' consent, but to undo the essence of that consent.

3. The ways forward? Contestability by design

Based on the indeterminacies identified above, the core argument of this paper is that TFAI risks prematurely collapsing the "superposition" of LOAC/IHL, especially in the contested areas of targeting decisions, forcing interpretive decisions to be taken at the design stage, instead of the operational stage, and thereby, displacing the discursive contestation that gives this area of law its perceived legitimacy notwithstanding its inherent indeterminacy. This paper believes that the appropriate role for TFAI should not be to guarantee automatic legal compliance, but to make breaches of LOAC/IHL detectable, sanctionable, and corrigible, and to surface the areas where human judgment is irreplaceable.[212]

This paper advocates the "contestability by design" feature, which requires an agentic AI system to flag, rather than resolve, existing legal indeterminacy. For example, when confronted with a targeting decision that engages the contested penumbra of distinction or proportionality, the AI system could present the human agent with at least two reasonable interpretations: a permissive reading and a restrictive humanitarian reading, together with the legal sources and reasoning

[211] One could compare this to the F-35 Joint Strike Fighters, which, although not an AI system, rely heavily on U.S.-controlled software, mission systems, and data links. This creates a long-term operational dependency on American channels for updates and maintenance. See Joseph Trevithick Rogoway Tyler, "How The F-35's Lack Of Spare Parts Became As Big A Threat As Enemy Missiles", *The War Zone*, 13 April 2023, available at: https://www.twz.com/how-the-f-35s-lack-of-spare-parts-became-as-big-a-threat-as-enemy-missiles.

[212] The current author thanks Dr Jon Truby for the idea of "contestability-by-design".

underpinning each. Military lawyers still need to select and justify a course of action, with the system logging both the interpretations adopted and the rationale for the chosen course. They would then bear moral, political, and legal responsibility for the interpretations adopted. This "contestability by design" corresponds to what Johan Rochel suggests as the user-based "panorama design".[213] As part of what has been labelled as the "participative turn in AI design", Rochel emphasises an approach that aims to surface "interpretative richness and normative disagreement", which could, in the context of LOAC/IHL, map out the qualification of conflicts, identify applicable norms, identify the relevant principles, and provide "a palette of potential interpretations of these principles (both in general and *in casu*)."[214] This design choice, Rochel observes, "reflects deeper normative commitments about how legal disagreement should be represented and who – the designer of the LLM or the user – bears responsibility for taking a decision considering those disagreements."[215] This reframing is particularly apposite for the indeterminate areas of LOAC/IHL, which this paper describes as a "superposition". The value of contestability by design lies in preventing a premature collapse of the "superposition" at the point of designing the AI-infrastructure, and preserving such "superposition" to the point of operational decision, where human participation in the interpretation, justification, and contestation of legal norms remains irreplaceable, and the sovereign authority to decide between different interpretations rests with a State's apparatus.

The proposal advanced here is, at root, an attempt to preserve LOAC/IHL's *ex post* justificatory function within an *ex ante* decisional architecture. By flagging rather than resolving indeterminacy, TFAI can support legal compliance without foreclosing the discursive space within which LOAC/IHL's *ex post* function operates. Although this suggestion would inevitably undermine TFAI's core value proposition of providing *ex ante* constraints on AI systems through automatic self-compliance, it reflects deeper normative commitments to LOAC/IHL's constitutive feature: the openness of argumentative space across interpretative communities. It also addresses a potential accountability gap in automatic legal interpretation by making the competing interpretations transparent and the underlying reasoning auditable. This paper acknowledges that the exact delineation between the determinate core and the indeterminate penumbra would require further study, which would hopefully involve military and humanitarian lawyers, computer scientists, technical experts, and the ICRC.

Conclusion

Hypothesising the Geneva Conventions and their Additional Protocols as AI-Guiding Treaties, this paper has examined how the specific characteristics of LOAC/IHL test the boundary conditions of the TFAI framework. Although legal

[213] J. Rochel, above note 36, pp. 26–27.
[214] *Ibid*., p. 26.
[215] *Ibid*.

positivism favours clarity and stability, advantageous for constraining the behaviour of States and their AI agents, it is problematic in the most controversial areas of LOAC/IHL. Returning to the "superposition" metaphor, this paper argues that the TFAI risks collapsing the desirable plurality of LOAC/IHL into either "apology" or "utopia". This paper further identifies three domain-specific challenges to the TFAI proposal as follows:

First, the "genuine agreement" thesis faces challenges from "strategic ambiguity". This paper's central analytical contribution is identifying "strategic ambiguity" as the meta-indeterminacy of LOAC/IHL, which explains why linguistic, legal, structural, and practice-based indeterminacies have persisted rather than been resolved. LOAC/IHL is a "hard case" for TFAI because it regulates the very instrument of sovereign power, where national interests necessarily diverge.

Second, the "objective determinacy" thesis works best at the field's uncontested core. Determinacy is not binary but exists on a spectrum, tracking the thickness of consensus between States. Rules at the determinate end, such as the prohibition on biological and chemical weapons, collective punishment, or the summary execution of POWs, reflect collective State interest and admit clear implementation by TFAI. Principles at the indeterminate end, precaution, distinction, proportionality, resist any reduction of life-and-death trade-offs to quantifiable standards. Demanding *ex ante* determinacy across LOAC/IHL is impossible and risks collapsing LOAC/IHL's "superposition" into State-to-State competition over *faits accomplis* and eroding the "strategic ambiguity" that underpins the field's very identity.

These challenges do not deny TFAI's value. Most LOAC/IHL violations throughout history have stemmed not from indeterminacies in legal interpretation but from human factors. A TFAI system implementing even the most permissive interpretation would still prevent a raging human commander from using autonomous weapons systems to execute civilians or POWs. Thus, TFAI proponents are indeed right that an imperfect *ex ante* constraint is still a significant improvement to the *status quo*. The qualification this paper adds, however, is that LOAC/IHL contains scenarios in which it does not, and is not intended to, provide a clear-cut answer. For instance, aerial strikes on military targets embedded among civilians remain contested with or without TFAI. The relevant comparison is therefore not between TFAI and no constraint, but between collapsing the "superposition" at the design stage and preserving it until the operational stage through the "contestability-by-design" feature.

Here, another TFAI's contribution to the field of LOAC/IHL is precisely to surface both the areas and stages of military decision-making in which human legal interpretation remains indispensable, precisely because it is contestable. Thus, by testing TFAI against perhaps the hardest case in international law, this paper does not intend to disqualify the TFAI proposal *per se* but rather to map out the terrain of

its general applicability. For instance, the concept of "strategic ambiguity" advanced here can surely be applied beyond LOAC/IHL to assess whether parties' ongoing disagreement prevented consensus, and which areas of international law did national interests between States sufficiently converge. The framework of three assumptions: (1) genuine agreement; (2) objective determinacy, and (3) neutral interpretation, provides future scholars with also a conceptual basis for implementing the TFAI proposal, possibly through *ex ante* and *de novo* legal specifications.[216]

All in all, this paper provides a more nuanced answer to Asimov's question as to whether autonomous agents could be reliably constrained by embedded legal rules than either his fiction or the TFAI proposal contemplates. The deepest test for international law in the age of AI is not whether the AI system can interpret and follow the law, but whether it can recognise the silences and nuances within the law that humans have deliberately preserved. Whether TFAI passes that test depends not on its technical sophistication but on the humility of its design.

[216] M. Maas and T. Olasunkanmi, above note 3, 94; on the role of technical standards, see, e.g., the role of ISO/IEC 42001:2023 standard, discussed in Jon Truby, "Global Governance of Transboundary Risks from Agentic AI: Enhancing International Standards for Auditing AI Management Systems to Comply with State Duties", in Philipp Hacker (ed), *Oxford Intersections: AI in Society*, 1st ed., Oxford University Press, Oxford, 2025.

Between Review, Transfer and Principle: The Limits of Article 36, the Arms Trade Treaty and the Martens Clause in AI Warfare

Dr. Anviksha Pachori[*]

ABSTRACT

The integration of artificial intelligence (AI) into weapon systems exposes serious accountability gaps in existing international legal frameworks. Article 36 of Additional Protocol I requires states to review new weapons for compliance with international humanitarian law, yet vague definitions and rapid AI integration blur their scope and weaken consistency. Doctrinal analysis reveals uneven state practice and limits standardized review processes amid growing autonomy and adaptability in weapons. The Arms Trade Treaty offers parallel safeguards but struggles with dual-use technologies and classification disputes. The Martens Clause provides interpretive guidance, yet gaps persist, underscoring the urgent need for clearer, adaptive frameworks governing AI-driven warfare

Keywords: Artificial Intelligence, Autonomous Weapons, Military, Technology, International Law

Introduction

"The genie is out of the bottle. The question is not whether AI will be used in warfare, but whether it will be governed." – Paul Scharre, Army of None.[1]

Artificial intelligence (AI) has emerged as a transformative force in contemporary warfare, posing profound challenges for the governance of weapons systems and the attribution of responsibility under international law. As states increasingly integrate AI into military decision-making, long-standing legal frameworks designed for the human-directed conduct of hostilities are being tested by systems capable of autonomous target selection, analysis, and action. Political leaders have already

[*]Associate Professor, G D Goenka, School of Law, New Delhi, India. Email- anvikshapachori@gmail.com. Orchid ID: https://orcid.org/0000-0003-2314-5598. I would like to express my sincere gratitude to Dr Jonathan Kwik, T.M.C. Asser Institute; Dr Samuel White, Centre for International Law, NUS; and Prof Sean Watts, Lieber Institute, for their invaluable feedback. Views are personal.

[1] Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, WW Norton & Company, New York and London, 2018.

proclaimed AI superiority as a determinant of global power,[2] prompting major military actors, including the United States and China, to accelerate investment in AI-integrated weapons systems.[3] This rapid militarization of AI has outpaced the development of corresponding legal and institutional mechanisms to ensure accountability. Notably, few nuclear-armed states have also pivoted towards integrating and developing advanced AI to bolster early warning, i.e., Intelligence, Surveillance and Reconnaissance (ISR), which could have consequences for strategic stability and the dominance of nuclear-powered countries.[4]

AI-integrated weapons are already altering the conduct of hostilities in ways that complicate compliance with international humanitarian law (IHL). In the Russia-Ukraine conflict, AI-enabled drones capable of autonomous navigation and target identification have reduced the immediacy of human control over targeting decisions, while Russia's deployment of systems such as the *Abzats*[5] anti-drone platform exemplifies the mounting reliance on algorithmic threat detection and response. This active deployment of ground robotics platforms on the battlefield of the Russia-Ukraine war has set a trend. More so in the Israel–Hamas conflict, systems including the Iron Dome[6] operate with a high degree of autonomy in detecting and intercepting threats, while AI-driven decision-support tools, such as *Habsora* and *Lavender* have contributed to target selection through algorithmic analysis of large datasets.[7] In the same line, South Asia has also witnessed its first AI-integrated drone-led war between India and Pakistan.[8] India launched drones like Shahpur-II with AI sensors which could identify strategic targets and adjust their trajectory before sending the coordinates to home systems.[9] In fact, the recent escalation in the US-Iran war illustrates that AI systems such as Project Maven do

---

[2] Patrick Caughill, "Vladimir Putin: Country that Leads in AI Development will be the Ruler of the World", *Futurism*, 30 November 2017, available at: https://futurism.com/vladimir-putin-country-that-leads-in-ai-development-will-be-the-ruler-of-the-world.

[3] Michael C. Horowitz, "Military AI Adoption is Outpacing Global Cooperation", *Council on Foreign Relations*, 11 February 2026, available at: https://www.cfr.org/articles/military-ai-adoption-is-outpacing-global-cooperation.

[4] Vladislav Chernavskikh, *Nuclear Weapons and Artificial Intelligence: Technological Promises and Practical Realties*, Background Paper, Stockholm International Peace Research Institute, September 2024, available at: https://www.sipri.org/sites/default/ files/2024-09/bp_2409_ai-nuclear.pdf.

[5] Christopher McFadden, "Abzats: Russia Deploys AI-powered Tiny Trucks to jam Ukrainian Drones", *Interesting Engineering*, 3 May 2024, available at: https://interesting engineering.com/innovation/russia-deploys-new-abzat-autonomous-drone-jamming-drone.

[6] "Israel-Hamas conflict: What is Iron Dome?", *BBC*, 10 October 2023, available at: https://www.bbc.com/news/articles/ck5xyz29e47o.

[7] Julius Kurek and Björn Laurin Kühn, *Habsora (הבשורה) and Lavender (לָבָנדֶר) Artificial Intelligence Systems*, Atlantic Forum, 19 September 2024, available at: https://www.atlantic-forum.com/our-views/habsora-and-lavender-artificial-intelligence-systems.

[8] Amoha Basrur, *The Use of Drones Marks a New Phase in India-Pakistan Hostilities*, Observer Research Foundation-Raisina Debates, 13 May 2025, available at: https://www. orfonline.org/expert-speak/the-use-of-drones-marks-a-new-phase-in-india-pakistan-hostilities.

[9] Harshul Sharma, "India vs Pakistan just Fought the First War Coded by Machine and Not Men", *PC Quest,* 13 May 2025, available at: https://www.pcquest.com/artificial-intelligence/india-vs-pakistan-just-fought-the-first-war-coded-by-machines-not-men-9062672.

not merely assist military operations but also restructure the temporal and epistemic conditions under which targeting decisions are made, thereby significantly increasing strike operations against Iran.[10] These developments raise pressing questions regarding the allocation of legal responsibility when AI systems influence or constrain human judgement, particularly in relation to the principles of distinction, proportionality, and precaution.

This article sets out to explore whether the existing international legal framework, *i.e.*, Article 36 of Additional Protocol I,[11] Arms Trade Treaty[12] (ATT) and Martens clause could meaningfully govern the integration of AI weapons. Article 36 presumes weapons whose behaviour can be assessed *ex ante*; AI-enabled systems evolve after deployment. Autonomous weapons systems, especially integrated with AI, should, like any other weapons systems, be subject to thorough legal review. This paper argues that reviews must extend beyond initial acquisition to encompass software, retraining processes, and operational modifications. One-time static certification is incompatible with evolving technologies. While this approach is already embedded in the International Committee for the Red Cross (ICRC) guidelines,[13] it should be treated as a legal requirement rather than just as "best practices". This is discussed in detail under section 2.

The ATT regulates discrete transfers of tangible arms. It assumes that arms transfers occur as discrete, identifiable events. But when a shipment leaves one jurisdiction to enter another, AI-enabled systems disrupt this model through what might be called continuous technological diffusion.[14] Thus, it demands epistemic certainty in a domain defined by technical uncertainty.[15] The more autonomous and adaptive the system, the less meaningful the Treaty's predictive calculus becomes.[16] In the AI context, compliance with the ATT may satisfy formal legality while leaving substantive humanitarian protection unchanged. That gap underscores the need for

---

[10] "Inside Palantir's Maven: The AI Tool Powering US Strikes in Iran", *The Times of India*, 27 April 2026, available at: https://timesofindia.indiatimes.com/defence/international/ middle-east-war-inside-palantirs-maven-the-ai-tool-powering-us-strikes-in-iran/article show/130183378.cms

[11] Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 36.

[12] Arms Trade Treaty, UNR 67/234B, 2 April 2013.

[13] International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, Geneva, August 2019, available at: https://www.icrc.org/sites/default/files/document/file_list/autonomy_artificial_intelligence_and_robotics.pdf.

[14] Paul Holtom and Mark Bromley, *Transit and Trans-shipment Controls in an Arms Trade Treaty*, Stockholm International Peace Research Institute, July 2011, available at: https://www.sipri.org/publications/2011/sipri-background-papers/transit-and-trans-shipment-controls-arms-trade-treaty.

[15] Tsvetelina Van Bentham, "Battlefield Uncertainties and the Use of AI", *International Law Studies*, Vol. 107, 2026, available at: https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=3129&context=ils.

[16] Lethal Autonomous Weapons Systems: Report of the Secretary-General, A/79/88, 1 July 2024, available at: https://docs.un.org/en/A/79/88.

design-oriented, continuous accountability mechanisms, which are advanced later in this paper. Crucially, this shift toward design-stage responsibility does not require the invention of new legal norms. It flows directly from existing IHL principles.

The cumulative analysis of this paper leads to a clear conclusion that Article 36 reviews alone cannot anticipate adaptive behaviour. Export controls cannot track intangible diffusion. The Martens Clause cannot substitute for enforceable rules. What remains is not a regulatory vacuum but a regulatory reorientation; compliance must migrate upstream, from procedure to architecture, from review to design. IHL-by-Design provides this reorientation. It translates enduring humanitarian principles into technical and institutional constraints capable of operating at the speed and complexity of AI-enabled warfare. Without such integration, existing frameworks risk devolving into formalities that legitimate deployment while leaving civilians exposed to algorithmically mediated harm. In short, if states remain responsible for the conduct of autonomous systems under international law, they must also be responsible for how those systems are built. The age of AI warfare, therefore, does not demand more norms and restrictions, but law embedded in its code and more transparency.

1. Autonomy of Weapon as a legal Disrupter

When IHL conceptualized weapons, it perceived weapons as a discrete tangible object whose effect and modes of operation were sufficient to have *ex-ante* legal evaluation. This assumption is embedded in Article 36 of the Additional Protocol I[17] which requires states to assess whether "a new method and means of warfare" would be prohibited under international laws. Based on article 36, it underpins core regulatory mechanism which presumes that a weapon's design, function and foreseeable effects can be assessed prior to deployment. AI enabled weapon systems fundamentally disrupt this premise. Unlike conventional weapons, their legal salient features do not reside primarily in their physical or kinetic performance, but in their complex computational decision-making processes that mediate perception, select, target, and engage with other objects and humans. Their behaviour is not solely determined at the point of manufacture but is continuously shaped by the data, updates, behaviour, observance and interactions.[18] As ICRC reports observed, autonomy in weapon systems shifts the locus of control from direct human decision making to algorithmic functions that operates in dynamic environments.[19] Consequently, the legal "object" of regulations shifts from a weapon thing to a decision-making process embedded within a broader system. This reconceptualization raises immediate challenges for IHL, which lacks doctrinal tools

[17] Additional Protocol I, above note 11, p. 4.
[18] Aparna Krishna Bhat, "The Evolution of AI: From Foundations to Future Prospects", *IEEE Computer Society*, 11 March 2025, available at: https://www.computer.org/ publications/tech-news/research/evolution-of-ai.
[19] ICRC Guidelines, above note 13, p. 4.

to regulate weapons whose most consequential features lie in the decision-making process rather than in their kinetic effects alone.

Under IHL, foreseeability is central to the lawful application of the principles of distinction and proportionality. Commanders are required to anticipate the effect of attacks, assess civilian harm and adjust conduct accordingly. On one hand, traditional weapons regulation assumes that such anticipation is grounded in weapons' known behaviour, on the other hand, AI-enabled systems, particularly those incorporating machine learning (ML), undermine this premise by introducing probabilistic and adaptive behaviour[20] that cannot be fully specified in advance. While proponents of Military AI often argue that such systems remain "predictable within parameters", this technical reassurance does not resolve the legal problem from an IHL perspective, a partial predictability is insufficient where legal responsibility hinges on reasonably foreseeable consequences. It is pertinent to note that increasing autonomy risks creating a "responsibility gap"[21] insofar as human operators lack sufficient understanding of how a system will behave in concrete operational settings. In the case of the ongoing contemporary wars, Russia and Ukraine are deploying loitering munitions such as drones that can patrol an area for extended periods and engage targets with varying degrees of autonomous guidance. The Iranian-designed *Shahed-136*, widely deployed by Russian forces, operates as a pre-programmed loitering munition capable of autonomously conducting surveillance and strike missions, akin to a long-range cruise missile, within an approximate 150 km radio-control range.[22] Similarly, Israel's defence forces have operationalized sophisticated AI systems such as Lavender, which automatically process vast amounts of surveillance and intelligence data to identify individuals and structures for targeting.[23] It maintains a large database of individuals linked to prospective militant groups, reportedly recommends effective targets using AI decision-support systems (AI-DSS), and aids in complex decision-making by processing large volumes of relevant information.

Critics argue that this AI-driven process blurs the distinction between "autonomy" and "intentions". Especially where human review is also limited or perfunctory, approaching this from a legal viewpoint, terms like "understanding" and "intentions" can be narrowly defined and have a huge space for implicit

---

[20] Major Vincenzo Gallitelli, "Artificial Intelligence-Enabled Military Decision-Making Process: The Forgotten Lessons on the Nature of War", *Journal of Advanced Military Studies*, Vol. 16, No. 2, 2025, p. 99.

[21] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Chistof Heyns, UN Doc. A/HRC/23/47, 9 April 2013.

[22] Bintari Rahmanita, "Specifications of Iran-Made Shahed-136 'Kamikaze' Drone", *TEMPO English*, 16 February 2025, available at: https://en.tempo.co/photo/126274/specifications-of-iran-made-shahed-136-kamikaze-drone.

[23] Emelie Andersin, "The Use of the 'Lavender' in Gaza the law of Targeting: AI-Decision Support Systems and Facial Recognition Technology", *Journal of International Humanitarian Legal Studies*, Vol. 16, 2025, p. 336.

assumptions.[24] IHL adopts a definition of intent that incorporates "awareness" that occurs in the ordinary course of events.[25] An act can be reckless or intentional, but will only exist when the system engages in or deals with conduct voluntarily triggered by an algorithm, leading to the attack of a civilian, not capable of distinguishing it from a military objective, leading to the death or injury of an innocent. In other words, Samuel and Brendon have defined it as an alignment problem in their paper, suggesting "a concept in the field of AI that describes the challenge of aligning the goals and behaviour of AI systems with the designer or operator's intention".[26] In this use case, AI doesn't merely assist human judgement by drastically narrowing the universe of targetable entities based on algorithmic inferences, it reconfigures the targeting process itself. When AI output significantly influences strike decisions, the legal burden shifts.

1.1. The Fragmentation of Agency and Dilution of Responsibility

The fundamental disruption introduced by AI-enabled weapons lies in the fragmentation of agency. At present, IHL and International Criminal Law (ICL) are premised on a relatively linear chain of responsibility where engineers build weapons, States deploy them, commanders decide, and operators execute. AWS disrupts this sequence by distributing decision-making across verticals from engineers, data trainers, procurement authorities, to algorithmic processes at machine speed. It not only obscures the question of responsibility but also blurs the line between the combatants and the role of civilians in a conflict.[27] Critics have argued, the delegation of critical functions to machines complicates the attribution of conduct and intent,[28] particularly where unlawful outcomes arise from emergent system behaviour rather than explicit human instruction. This does not entirely dismiss the State's responsibility under the law of State responsibility; however, it significantly weakens the practical mechanism of accountability, especially in the absence of transparency regarding system design and operational parameters.

AI-enabled weapons systems also disrupt the temporal assumptions underpinning weapons regulation. Article 36 reviews are typically conducted at a discrete point prior to adoption or deployment on the assumption that the weapon's characteristics will remain materially constant. AI systems, however, evolve through software updates, retaining and contextual learning often after deployment. As the

---

[24] International Criminal Tribunal for former Yugoslavia, Prosecutor v. Zejnil Delalic, Case No. IT-96-21-T, Judgement, 1 November 1996.

[25] Magdelena Pacholska, "Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective", *Israel Law Review*, Vol. 56, No. 1, p. 3. 2022, available at: https://doi.org/10.1017/S0021223722000188.

[26] Samuel Hartridge and Brendan Walker-Munro, "Autonomous Weapons Systems and the AI Alignment Problem", *Journal of International Humanitarian Legal Studies*, Vol. 16, 2025, p. 38.

[27] Andreas Wenger and Simon J.A. Mason, "The Civilianization of Armed Conflict: Trends and Implications", *International Review of the Red Cross*, Vol. 90, 2012, p. 872.

[28] Michael N. Schmitt and Jeffery S. Thurnher, "Out of the loop: Autonomous Weapon Systems and Law of Armed Conflict", *Harvard National Security Journal*, Vol. 4, No. 2, 2013, p. 231.

ICRC emphasized that systems capable of modifying their behaviour raise questions about whether legal review should be understood as a continuous obligation rather than a one-off procedural step.[29] The temporal fluidity challenges the binary distinction between lawful and unlawful weapons. A system that appears compliant at the design stage may behave differently in a specific operational environment, raising questions about when legal responsibility crystallizes and whether compliance can be meaningfully assessed at a single time.[30] This misalignment does not imply a legal vacuum, but it does necessitate a shift in regulatory focus from static weapon assessments to process-based, continuous and design-oriented regulation. Without such a shift, mechanisms like Article 36 risk functioning as symbolic safeguards rather than effective tools of accountability.

2. Review of Weapons under Article 36 and State Practices

It is an established norm by the International Court of Justice (ICJ) in its *Advisory Opinion on the Legality and Threat or Use of Nuclear Weapons* that, "IHL permeates the entire law of armed conflict and applies to all forms of warfare and to all kinds of weapons, those of the past, those of the present and those of the future."[31] The ICJ also noted that this prohibition ensures that "States do not have unlimited freedom of choice of means in the weapons they use,"[32] and declared the rule as one of the cardinal principles contained in the texts constituting the fabric of IHL. Legal reviews are frequently portrayed as a foundation of the domestic implementation of IHL. Although Article 36 imposes a binding obligation on States Parties to determine the legality of new weapons, means, and methods of warfare, the practical effectiveness of this mechanism remains uneven and heavily dependent on national discretion.[33] In the absence of uniform international standards, independent oversight, or reporting obligations, Article 36 reviews often risk becoming opaque procedural exercises rather than meaningful accountability mechanisms.[34] Only a limited number of States, including the Netherlands, Sweden, Switzerland, and Australia, have publicly disclosed aspects of their weapons review procedures, including their potential application to autonomous or AI-enabled weapon

[29] Kathleen Lawand, *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*, ICRC, Geneva, 2006.
[30] Jonathan Kwik, "Iterative Assessment for Military Artificial Intelligence (AI) Systems" in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, T.M.C. Asser Press, The Hague, 2026, available at: https://doi.org/10.1007/978-94-6265-759-5.
[31] International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons,* Advisory Opinion*, ICJ Reports* 1996, 8 July 1996, para. 86.
[32] *Ibid.*, para 78.
[33] Additional Protocol I, above note 11, p. 4; K. Lawand, above note 29, p. 6.
[34] Dan Saxon (ed.), *International Humanitarian Law and the Changing Technology of War*, Martinus Nijhoff, Leiden, 2013.

systems.[35] Such disclosures, however, are largely confined to general descriptions of institutional processes and legal criteria, rather than substantive assessments of specific systems or technologies. Consequently, while transparency in the conduct of legal reviews is frequently invoked as a confidence-building measure, the restricted visibility into review methodologies, assumptions, and outcomes significantly constrains their capacity to ensure effective compliance with treaty and customary IHL governing the means and methods of warfare,[36] particularly in the context of increasingly autonomous and adaptive weapons systems.

Article 36 of Additional Protocol I is often portrayed as the foundation of lawful weapons regulation in emerging technological contexts. It obliges states to determine, during the "research, development, acquisition or adoption" of a new weapon, whether its use should be prohibited under international law. This obligation is preventive in nature, seeking to internalize compliance with IHL before harm occurs. However, this preventive logic rests on a critical but under-examined assumption that the weapon under review is stable and knowable characteristics to permit a meaningful legal assessment *ex ante*.[37] AI-enabled weapons directly undermine this assumption. Unlike conventional weapons, whose operational parameters remain largely fixed after deployment, AI-driven systems may alter behaviour through learning, environmental interaction or software modification. As a result, invoking the concept of "meaningful human control" cannot be determined in AWS as a range of potential future behaviours many of which cannot be fully anticipated at the point of review. The problem, therefore, is not that Article 36 is ignored by many States but that it is asked to perform a function for which it was never designed. ICRC has also acknowledged that meaningful human control functions more as a policy aspiration than a legal standard, offering limited guidance for concrete review decisions.[38]

Perhaps the most significant limitation of Article 36 in the context of AI lies in temporal rigidity. Weapons reviews are typically conducted once, prior to the

---

[35] Netherlands Ministry of Defence, Advisory Report on Autonomous Weapon Systems (2016); Also See, "Statement by Sweden at the CCW GGE on Lethal Autonomous Weapons (LAWS)", *UNODA*, 21-25 September 2020, available at: https://documents.unoda.org/wp-content/uploads/2020/09/200921-Anforande-LAWS-human-control_.pdf; A "compliance -based" approach to Autonomous Weapon System, CCW/GGE.1/2017/WP.9, 10 November 2017, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional _Weapons_-_Group_of_Governmental_Experts_(2017)/2017_GGEonLAWS_WP9_ Switzerland.pdf; The Australian Article 36 Review Process, CCW/GGE.2/2018/WP.6, 30 August 2018, available at: https://docs-library.unoda.org/Convention_on_Certain_ Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/2018_GGE%2 BLAWS_August_Working%2Bpaper_Australia.pdf.

[36] International Committee of the Red Cross, "Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons", 1 September 2017, available at: https://www.icrc.org/en/publication/4283-autonomous-weapons-systems; Rebecca Crootof, "The Killer Robots Are Here: Legal and Policy Implications", *Cardozo Law Review*, Vol. 36, 2015, p. 1837.

[37] Additional Protocol I, above note 11, p. 4.

[38] ICRC, above note 13, p. 8.

deployment and rarely revisited unless a weapon undergoes substantial modification. On the other hand, AI-embedded systems are evolving continuously. This evolution may materially affect how the system identifies targets, prioritizes threats, detects or executes force. The legal consequences of this are stark i.e. when a weapon deemed lawful at time T1 may behave illegally at time T2; the weapon being properly reviewed but the illegality not being foreseeable. The critics have argued that treating legality as a static attribute of a weapon is increasingly untenable[39] where the decisive features of the system are dynamic.

Without a doctrine of continuous review or post-deployment accountability, Article 36 risks becoming disconnected from operational realities of AI-governed weapons. The shortcomings of Article 36 in relation to AWS should not be understood as failures of implementation alone; rather, they expose a structural inadequacy rooted in the provisional reliance on stability, foreseeability and one-time review temporally fixed assessments. These assumptions are incompatible with technologies that re-distribute decision-making authority across machines, humans and time. But this does not render Article 36 irrelevant; rather, it necessitates an evolution in how it is understood and applied. It just requires an evolution of its own understanding and application. In the further sections, the paper captures the reconceptualization of Article 36's function from one-time legality to a continuous, design-oriented responsibility approach.

2.1. Procedural legitimacy and constraint under Article 36

State practice relating to Article 36 of Additional Protocol I demonstrates a striking paradox where, on one hand, the obligation to conduct national weapon reviews has achieved near-universal normative acceptance, but on the other hand, with AWS, its practical operation remains fragmented, opaque and uneven. The cumulative evidence suggests that Article 36 functions less as a robust constraint on unlawful or destabilizing weapons and more as a procedural assurance mechanism whose effectiveness depends heavily on domestic capabilities to perform the evaluation and the willingness of the nation. As an aggravating factor the complexities of AWS are growing each day, the more complex the architecture of the system the harder and more expensive it becomes to perform separate tests and evaluations.[40] The complexity of the environment in which the system is intended to operate is another important variable that further muddles and increases the cost of testing and evaluation which in turn is directly linked with the budget of the country and its capacity to sustain that cost of further analysis. The current assessment under Article 36 is neither practical nor financially prudent for AI-enabled AWS using the existing

[39] Marco Sassoli, *International Humanitarian law: Rules, Controversies and Solutions,* 2nd ed., Edgar Elgar Publishing, Northampton, 2019.
[40] "Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects", *International Review of the Red Cross*, 28 March 2014, available at: https://www.icrc.org/sites/default/files/document/file_list/4221-002-autonomous-weapons-systems-full-report.pdf.

testing and evaluation method.[41] It is paramount to develop new methodologies for testing and evaluating new weapons. Some states, such as the United States, are currently conducting important research and development efforts to address the issue. For AI-enabled weapons, a system should report on its learning and modifications throughout its lifecycle, allowing the user to obtain a new review mechanism.[42] At the moment, the creation of such "self-aware" or "self-explainable" AWS remains at the stage of research itself.

2.2. State Practices Developed under Article 36

Although at the level of principle, the contestation on Article 36 is minimal. A large majority of participating States affirm that they are legally bound to review new weapons, means and methods of warfare. Few countries treat the obligation as reflective of customary international law.[43] Institutionally, State practice reveals three distinct models of practices. First groups comprise of countries such as Australia, the Netherlands, Germany, Norway among others who have established relatively permanent structured review mechanisms that integrate legal, technical, operational and sometimes medical expertise.[44] These systems tend to embed legal advice early in the procurement and development cycle and in some cases life-cycle assessment. Such institutionalization enhances professionalism and predictability of the weapons in general. The second group relies on *ad hoc* or procurement-driven assessments, conducting reviews only when acquisition decisions arise.[45] Although these countries are formally compliant, such models risk becoming reactive rather than preventive. The purpose of Article 36 being an *ex-ante* safeguard is to scrutinize legality during research and development yet this episodic review may simply rationalize decisions already made for strategic reasons.[46] Finally, a third model

[41] 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 Weapon Reviews", *International Review of the Red Cross,* Vol. 94, No. 886, 2012, p. 483.
[42] Vincent Boulanin, Laura Bruun and Netta Goussac, *Autonomous Weapons Systems and International Humanitarian Law, Identifying Limits and the Required Type and Degree of Human-Machine Interaction*, Stockholm International Peace Research Institute, June 2021, available at: https://www.sipri.org/sites/default/files/2021-06/2106_aws_and_ihl_ 0.pdf.
[43] 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*, Vol. 88, No. 864, 2006.
[44] *Ibid.*, p. 931.
[45] Rosie Cavdarski, Lauren Sanders & Rain Liivoja, "Legal Reviews of Autonomous Weapon Systems: Report on Submissions to the GGE on LAWS (version 2.0)", *Law and the Future of War Research Group*, The University of Queensland, Australia, 2023, available at: https://doi.org/10.14264/1c0275b.
[46] Vincent Boulanin, *Implementing Article 36 Weapon Reviews in the Light of Increasing Autonomy in Weapon Systems*, SIPRI Insights on Peace and Security, November 2015, available at: https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf.

provides little or no detail about their review mechanism, despite affirming adherence to the obligation.[47]

The effectiveness of Article 36 is strongly based on institutional design, procedural standards and formalized mechanisms. The prime purpose of review is to assess compatibility with core principles of IHL i.e. distinction, proportionality, precaution, necessity of attack and other treaty-based restrictions. The first group of states have embedded review with these set standards. Australia provides one of the most detailed and structured examples. All weapons intended for operational use must undergo testing. The review produces one of three determinations such as "cleared", "cleared subject to conditions" or "not cleared for operational use".[48] Australia has also introduced an annual national course to certify legal advisers responsible for conducting reviews. This model treats review as a preventive governance rather than a *post hoc* justification. The Netherlands adopts an equally institutionalized approach; its review process is explicitly multidisciplinary, involving legal, technical, medical and environmental expertise, culminating in final ministerial approval.[49] Such integration reflects an understanding that compliance cannot be reduced to purely doctrinal analysis but must incorporate operational realities and ethical risk assessment.

Germany emphasizes structural independence and procedural safeguards; their reviews are formalized through Central Service regulations and conducted by qualified legal, technical and operational experts who maintain hierarchical independence from both developers and end users.[50] It also recognizes that programming modifications that alter the system's or ML's functioning require renewed review. By foregrounding independence and re-review, Germany addresses the conflict of static review and the problem of one-time approval. Other states have adopted a more declaratory posture. The UK characterizes legal reviews as "gatekeepers", ensuring that weapons capable of lawful use reach the battlefield, yet provides comparatively little detail on procedures or outcomes.[51] Many states differentiate between sharing general processes and withholding substantive findings;

---

[47] 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*, Vol. 28, No. 2, p. 285.

[48] *Ibid.*

[49] Vincent Boulanin and Maaike Verbruggen, *SIPRI Compendium on Article 36 Reviews*, Stockholm International Peace Research Institute, December 2017, available at: https://www.sipri.org/sites/default/files/201712/sipri_bp_1712_article_36_compendium_2017.pdf.

[50] Federal Foreign Office, "German Commentary on "Operationalizing all Eleven Guiding Principles at a National Level as Requested by the Chair of the 2020 Group of Governmental Experts (GGE) on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS) within the Convention on Certain Conventional Weapons (CCW)", *UNODA*, 24 June 2020, https://documents.unoda.org/wp-content/uploads/2020/07/20200626-Germany.pdf.

[51] "United Kingdom: Item 5 – Topic 5: Legal Reviews", *UNODA*, 9 March 2023, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_%282023%29/UK_Intervention_Item_5_Topic_5_Legal_Reviews.pdf.

they neither share the criteria nor reject the decision post-review, which could be publicly verifiable. Nineteen European Union member states and the ICRC support real information sharing and the exchange of best practices.[52] Argentina, among others, have argued that transparency is necessary to reduce disparities between technologically advanced States and procurers with limited evaluative capacity. A few countries, like China and Russia, have also resisted stronger transparency obligations.[53]

2.2.1. State Practice and AWS

These structural limitations are magnified in the case of AWS. Unlike conventional weapons, autonomous systems challenge the traditional logic of review because their behaviour may evolve through software updates, data inputs, machine learning, or contextual deployment rendering one-time static review and approval as inadequate. Thirteen States recognize this, a growing number of States emphasize the need for life-cycle review, reliability testing, and reassessment following modifications. Some explicitly argue that changes to programming or learning functions should trigger a renewed review.[54] Norway highlights that programming changes the system's functioning and necessitates reassessment.[55] Similarly, Germany treats algorithm or behavioural changes as triggers for fresh legal scrutiny. Yet Pakistan contends that the nature of AI makes continuous review effectively impossible.[56] This disagreement underscores a deeper tension that Article 36 was designed for discrete weapon objects, whereas AWS demands continuous, system-level oversight. Without procedural adaptation, review risks becoming obsolete precisely where scrutiny is most needed. In short, State practice demonstrates that Article 36 has succeeded in entrenching a global expectation that weapons be reviewed but has not yet succeeded in ensuring that such reviews function as genuine barriers to unlawful or destabilizing systems. The norm is procedurally entrenched but substantively fragile. AWS development exposes this fragility and makes clear that, unless review mechanisms evolve toward greater consistency and accountability, Article 36 may legitimize rather than regulate the next generation of weapons.

---

[52] "Convention on Certain Conventional Weapons – EU Statement on Lethal Autonomous Weapons Systems (LAWS)", *European Union External Action*, 13 November 2017, available at: https://www.eeas.europa.eu/node/35508_fr.

[53] Approaches of the Russian Federation to the Issue of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems: Submitted by the Russian Federation, CCW/GGE.1/2024/WP.2, 1 March 2024.

[54] Thirteen states and the ICRC consider that modifications to AWS require renewed review.

[55] "Submission by Norway on Lethal Autonomous Weapons Systems", *UNODA*, 24 May 2024, available at: https://docs-library.unoda.org/General_Assembly_First_Committee_-Seventy-Ninth_session_%282024%29/78-241-Norway-EN.pdf.

[56] "Pakistan", *Automated Decision Search*, available at: https://automatedresearch.org/news/state_position/pakistan/.

3. The Arms Trade Treaty and the Disaggregation of Responsibility

The ATT is only a decade-old treaty, albeit with 20 years in the making.[57] It is presented as a complementary safeguard to IHL compliance, operating upstream of battlefield use by regulating the international transfer of conventional arms. The safeguards and/or safety mechanisms address the transfer of weapons to another country. Thus, the most ATT can do, in its most ideal form, is stop a country with weapons from giving them to another country which is perhaps aiming to violate human rights with such weapons. In addition, the international arms trade accounts for only a fraction of total military expenditure globally.[58] However, if one were to look at conflicts, for instance, the Yemen Conflict and Israel-Hamas war, the importance of optimal application of ATT resurges. Saudi Arabia used weapons on schools, hospitals, markets and weddings, even contributing to famine-like situations; while Israel has been using air-dropped bombs and artillery, also destroying medical facilities and refugee camps.[59] Both countries are importers of weapons and indulge in the arms trade to facilitate their operations. Both imported about 70% of their arms from the United States of America.[60] This raises the question of the supplier's responsibility for these apparent human rights violations.

The role of AI and the rise of AWS further complicate the question of responsibility. The first documented use of AWS was in 2021 during the Libyan War.[61] So, even though the research has been in place for such technology and automation since the 1950s,[62] the use is fairly recent but steady for both AI and automation in weaponry.[63] Now, AI and automation may be complementary but have distinct roles to play in terms of responsibility.

---

[57] Peter Woolcott, "Arms Trade Treaty: Introductory Note", *Audiovisual Library of International Law*, 2 April 2013, https://legal.un.org/avl/ha/att/att.html.
[58] Xiao Liang *et al.*, *Trends in World Military Expenditure, 2024*, Stockholm International Peace Research Institute, April 2025, available at: https://www.sipri.org/sites/default/ files/2025-04/2504_fs_milex_2024.pdf.
[59] "Press Release: New Study Shows Israel Air-dropped 2000lb Bombs Within Lethal and Damage Ranges of Hospitals in Gaza", *FXB Center for Health and Human Rights at Harvard University*, 10 October 2024, available at: https://fxb.harvard.edu/blog/2024/10/10/new-study-shows-israel-air-dropped-2000lb-bombs-within-lethal-and-damage-ranges-of-hospitals-in-gaza/.
[60] "US Arms Exports", *Campaign Against Arms Trade*, 22 January 2026, available at: https://caat.org.uk/data/countries/united-states/us-arms-exports/.
[61] Majumdar Roy Choudhury *et al.*, Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council, S/2021/229, 8 March 2021, available at: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N21/037/72/PDF/N2103772.pdf?OpenElement.
[62] Goddy Uwa Osimen, John Anegbode and Temidayo Oladipo, "Autonomous Warfare: Exploring the Consequences of AI- Powered Weapons System on Global Peace and Security", *Power Tech Journal*, Vol. 49, No. 2, 2025, p. 1458.
[63] *Ibid.*

3.1. Risk Knowledge in a world of software, components, and updates

The first structural difficulty concerns the treaty's object of the regulation. ATT applies to the "conventional weapons" listed under article 2(1), categories historically tied to physical platforms and munitions.[64] AI enabled capabilities, by contrast are often embedded in software updates, training datasets, algorithms and cloud-based services rather than in tangible hardware. This raises an important question of "what exactly is being transferred?" If the decisive targeting capabilities of an AWS reside in codes rather than in the airframe or munition, export control that only focuses on the physical system would regulate the shell, not the substance. As the Stockholm International Peace Research Institute (SIPRI) has observed, emerging military AI technologies frequently fall outside traditional arms classifications[65] because they resemble dual-use software rather than weapons per se.

The very fundamental obligation of the ATT is to assess the intended use of the weapons before green-lighting the trade. The ATT provides for this obligation within its text under Article 6 and 7. Article 6 places an obligation on the states to not authorise or stop the transfer if it has "knowledge" of any untoward use such as to commit genocide, crimes against humanity, grave breaches of the Geneva Conventions, attacks directed against civilians and/or other war crimes.[66] Article 7, on the other hand, provides for the overriding risk standard.[67] States are obligated, under Article 7, to assess the potential that the arms/items would undermine peace and security, and/or be used to commit or facilitate serious violations of international humanitarian law and human rights law, and acts of terrorism and transnational organized crime to which the exporting state is a party.[68] When this risk outweighs other considerations, the exporting State must refuse to transfer.

Both articles provide for the minimum knowledge or justified belief required before the law assigns responsibility or triggers a legal consequence. Both determine when a State must refuse to authorise an arms transfer. This is the epistemic threshold of ATT. This requirement to assess risk and deny when there's knowledge of untoward consequences at the time of transfer, on the surface, seems good enough. But the advent and involvement of AI and automation pose a significant issue: how do you evaluate the risks which may evolve across time and actors?

The ATT, in its existing text, presumes identifiable weapons and the time of transfer, assessable future risk and concentrated responsibility, which perhaps works

---

[64] Arms Trade Treaty, UNR 67/234B, 2 April 2013, Art. 2 (1).
[65] Yasmin Afina and Giacomo Persi Paoli, "Governance of Artificial Intelligence in the Military Domain: A Multi-Stakeholder Perspective on Priority Areas", *UNIDIR*, 5 September 2024, available at: https://unidir.org/publication/governance-of-artificial-intelligence-in-the-military-domain-a-multi-stakeholder-perspective-on-priority-areas/.
[66] Arms Trade Treaty, UNR 67/234B, 2 April 2013, Art. 6.
[67] Arms Trade Treaty, UNR 67/234B, 2 April 2013, Art. 7.
[68] *Ibid.*

well for conventional arms. But with software systems, iterative updates, modular components and learning systems, the assumptions upholding the effectiveness of "risk knowledge" become inconsistent at best, and ineffective at worst. Risks, now, are not assessable *ex ante*. With AI and automation, the risks may evolve after the export is already done, knowledge may not even be available, harm may emerge through integration, causation may not be linear, and most importantly, the attribution of responsibility may be fragmented at best, and non-existent at worst.

To explain why AI supply chains defeat this linear risk assessment, we must examine what it assumes and how AI supply chains operate. Linear risk assessment is a simple approach in which causation flows in a straight line and the risk assessment occurs at a single identifiable decision point. It assumes that there is an identifiable object, stable functionality, single transfer of technology, traceable causation and, again, centralised responsibility. AI supply chains, however, are update-driven not static, modular not monolithic, and recursive not stagnant. Firstly, AI supply chains being update driven implies they get updated, undergo training and retraining, and adjust and improve over time. Secondly, AI supply chains being modular imply that they consist of separate modules rather than a single technology. Thus, each module can be developed, exported and classified differently.[69] Recursive AI supply chains imply that they are adaptable and can incorporate feedback from deployment into future behaviour.

The update-driven, modular, and recursive nature of AI supply chains undermines temporal stability, disrupts responsibility, and undermines causal traceability. Risks, then, cannot be reasonably foreseen and the knowledge is abstract, if it is even available. This is what the scholars refer to as "temporal misalignment"[70] between authorisation and operation.[71]

3.2. Dual-use AI and the collapse of the conventional arms boundary

Conventional arms control law concerns itself with tangible, military-design objects with the primary object or operation being killing/injuring.[72] Non-arms refers to civilian goods, services and/or technologies. This differentiation or boundary collapses when AI has dual-use. Dual-use AI refers to artificial intelligence systems which can slide between benign and harmful uses depending on context, integration,

---

[69] Aspen Hopkins *et al.*, "On Deployment: AI Supply Chains", *Social Science Research Network*, 3 April 2013, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_ id=4789403.
[70] Wenjing Lou *et al.*, "Temporal Misalignment Attacks against Multimodal Perception in Autonomous Driving", *ArXiv*, 1 October 2025, available at: https://arxiv.org/html/ 2507.09095v2.
[71] Sze Hong Lam, "Will Treaty Following AI Collapse into the dichotomy between 'Apology' and 'Utopia'"? *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare*, 2026, pp.48-81.
[72] Frank Sauer, "Autonomy in Weapons Systems and the Struggle for Regulation", *Center for International Governance Innovation*, 28 November 2022, available at: https://www.cigionline.org/articles/autonomy-in-weapons-systems-and-the-struggle-for -regulation/.

and intent without any change to their constitution.[73] Thus, being able to be used for both legitimate civilian purposes and for military, security or coercive applications. Examples of dual-use AI include facial recognition technologies and predictive analytics or logistics. At present, there is a great number of potential applications of biometrics used in the military domain; it can be used throughout the full range of operations, including combat, security, and peace support operations. In general, biometrics allows verification of personal identity and identification of enemy threats with a higher degree of confidence when personal enters military facilities.[74] Military power is no longer concentrated in tangible arms, but spread across algorithms, data and upgrades that could become non-traditional weapons.

Algorithms are capable of being used as weapons. Algorithms can be used to select targets, prioritise threats, optimise strike decisions and/or automate surveillance.[75] Algorithms may not be traditionally weapons but can be operated just like one. Secondly, data enables identification, tracking, prediction and classification of populations. Data also is not a traditional weapon but can transform people into precise targets, making violence administratively scalable.[76] Thirdly, to understand upgrades as weapons requires a bit more understanding. Upgrades are continuous, service-based, framed as support and innovation and delivered across borders digitally. No transfer tangibly occurs, no weapon changes hands, and no review is triggered with each update. However, with each upgrade, the lethality increases. An upgrade in itself is not a weapon. Upgrades of AI systems are sites where lethal capacity is introduced, intensified, or transformed, outside the object-based logic of arms control law. Moving from ontology to governance now. Modularisation refers to breaking down a system into separable components, each developed, sold and transferred independently.[77] Responsibility and governance then become convoluted and invisible. Modularisation enables legal evasion. AI distributes capability across modules. There is no single transfer to trigger arms laws. States and/or entities

[73] Jayshree Pandya, "The Dual-Use Dilemma of Artificial Intelligence", *Forbes*, 7 January 2019, available at: https://www.forbes.com/sites/cognitiveworld/2019/01/07/the-dual-use-dilemma-of-artificial-intelligence/.

[74] Marten Zwanenburg, "Know They Enemy: The Use of Biometrics in Military Operations and International Humanitarian Law", *International Law Studies Stockton Centre for International Law*, Vol. 97, 2021, US Naval War College, available at: https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=2986&context=ils.

[75] Anthony Downey, "The Alibi of AI: Algorithmic Models of Automated Killing", *Digital War*, Vol. 6, No. 9, 2025, available at: https://link.springer.com/article/10.1057/s42984-025-00105-7.

[76] Lucy Suchman *et al.*, "Tracking and Targeting: Sociotechnologies of (In)security", *Science Technology and Human Values Journal*, Vol. 42, No. 6, 2021 available at: https://journals.sagepub.com/doi/full/10.1177/0162243917731524?casa_token=ayRc299r33cAAAAA%3A6Jc7GYyyBvyJISrbkCsxjUnMhJmYrMQioCsajxzr-hXG9ZKJ-Y1BHOlIm7v-rUGV5qwdsvqaOvKm.

[77] Paolo Trucco *et al.*, "Modularisation as a System Life Cycle Management Strategy: Drovers, Barriers, Mechanisms and Impacts", *International Journal of Engineering Business Management*, Vol. 11, 2019, available at: https://journals.sagepub.com/doi/ epub/10.1177/1847979018825041.

engage in licensing agreements, instead of direct transfers. Modularisation, then, does not violate but bypasses the existing laws.[78]

The IBM case perfectly encapsulates the complexity of dual-use AI, modularisation and legal evasion. IBM's punch-card tabulation systems, supplied to and serviced within Nazi Germany, were not weapons in any conventional sense.[79] They were data-processing machines, commercially marketed, ostensibly administrative, and entirely compatible with ordinary civilian use. Yet they became infrastructural instruments of persecution.[80] What rendered them consequential was not their material form, but their integration into a broader architecture of state violence. The harm did not arise from a single transfer event. It was sustained through modular components: machines, punch cards, technical servicing, updates, and maintenance. Each of these could be framed as commercially routine. Responsibility was therefore distributed across functions rather than concentrated in a discrete act of arming. No rifle crossed a border; no bomb was supplied.

This structural dynamic mirrors the contemporary problem of dual-use AI. Contemporary examples include Amesys selling their internet inspection system to Gaddafi's Libya, which was used to monitor dissidents, journalists and activists.[81] Executives of Amesys and Nexa were indicted for complicity in torture.[82] However, the transfer of technology never triggered any ATT compliance. Algorithms, datasets, and software upgrades are rarely lethal in isolation. Their significance lies in their embedding, recombination, and iterative enhancement within operational systems. Modularisation allows each component to appear neutral, while its cumulative integration materially reshapes the capacity for surveillance, targeting, or repression. Legal scrutiny that focuses only on finished weapons or discrete export events risks overlooking this distributed architecture of enablement.

3.3. Export control without accountability

To revisit the case of Amesys and Libya, Amesys is a technology company based in France. Amesys supplied its technology to Libya and Egypt. Even though there was ATT at the time, the technological transfer was not scrutinised. ATT compliance, as it is, is procedural, not substantive. The authorisation that is given under ATT provides for *ex ante* obligations for the exporting state which are, as discussed, insufficient in light of AI systems. Public international law generally also tolerates

---

[78] Neil Cooper, "What's the point of arms transfer control?", *Contemporary Security Policy*, Vol. 27, No. 1, 2006, p. 118.
[79] Harry Murphy, "Dealing with the Devil: The Triumph and Tragedy of IBM's Business with the Third Reich", *The History Teacher*, Vol. 53, No. 1, 2019, p. 171.
[80] *Ibid.*
[81] MEE Staff, "French Spy-tech Firms Accused of 'Complicity in Torture' in Libya and Egypt", *Middle East Eye*, 22 June 2021, available at: https://www.middleeasteye.net/news/ egypt-libya-french-spy-tech-firms-accused-complicity-torture.
[82] *Ibid.*

due diligence standards not strict liability.[83] Accountability with AI systems and autonomous weapons is distributed and harder to uphold. It is distributed across technical, legal and temporal functions, a problem of too many chefs. With such diffusion of responsibility among developers, exporters and states, the law is not violated but the legality is restructured. Traditional attribution doctrines such as state responsibility, complicity, aiding and abetting struggle with cumulative causation. This becomes apparent in the IBM case as well. Even though the involvement was documented and has never been denied, the company was never held accountable.[84] To draw on Ulrich Beck's organised irresponsibility, ignorance is not the absence of knowledge but a systemically generated condition.[85]

The ATT focuses on conventional arms, the assessment of foreseeable risk, and the denial of supply if there is an overriding risk. With AI systems, three problems arise: firstly, compliance under ATT is event-based, whereas harm from AI systems is process-based; secondly, ethical risks may be known but legally insufficient to deny supply under Article 7; and thirdly, modularisation enables formal compliance while functionally evading control. These features allow ATT compliance to operate as a procedural rather than a mechanism of substantive containment. Once an export authorisation is granted, the Treaty does not impose a continuing duty to reassess evolving system behaviour, even where subsequent software updates, retraining, or system integration materially alter the risk profile.[86] The regulatory gaze remains fixed on the moment of transfer, while the locus of harm shifts to post-export modification and use. Compliance is thus temporally satisfied even as risk migrates.

In the context of AI-enabled systems, risks associated with autonomy, algorithmic bias, or downstream repurposing are often foreseeable in an ethical sense but remain insufficiently determinate to meet the Treaty's legal threshold.[87] This creates a space in which States may acknowledge serious concerns while still lawfully authorising transfers, thereby fostering ethical awareness without legal consequences. Modularisation further entrenches this gap. By disaggregating capability across software, data, hardware, and ongoing services, AI supply chains

---

[83] Mark Taylor *et al.*, "Human Rights Due Diligence: The Role of States", *Human Rights in Business*, December 2012, available at: https://humanrightsinbusiness.eu/wp-content/uploads/2015/05/De-Schutter-et-al.-Human-Rights-Due-Diligence-The-Role-of-States.pdf.

[84] H. Murphy, above note 79, p. 19.

[85] Maria Galantino, "Organised Irresponsibility in the Post-Truth Era: Beck's Legacy in Today's World at Risk", *Italian Sociological Review*, Vol. 12, No. 8S, 2022, available at: https://www.proquest.com/openview/a1f67cfd911f2e2741b7e59304d145a1/1?pq-origsite=gscholar&cbl=1796337.

[86] Paul Beijer, *Taking Stock of the Arms Trade Treaty: Application of the Risk-assessment Criteria*, Stockholm International Peace Research Institute, August 2021, available at: https://www.sipri.org/publications/2021/policy-reports/taking-stock-arms-trade-treaty-application-risk-assessment-criteria.

[87] "A Hazard to Human Right: Autonomous Weapons Systems and Digital Decision-Making", *Human Rights Watch*, 28 April 2025, available at: https://www.hrw.org/report/ 2025/04/28/a-hazard-to-human-rights/autonomous-weapons-systems-and-digital-decision-making.

allow each transaction to be assessed in isolation, none of which appears to trigger the prohibitive logic of Article 6 or 7.[88] The cumulative system, however, may enable forms of surveillance, targeting, or repression that the Treaty was designed to prevent. Responsibility is thereby diffused across actors and transactions, while no single authorisation bears the weight of the eventual harm. In this configuration, export control remains formally intact but normatively hollow. The ATT succeeds in regulating compliance at the level of authorisation, yet fails to ensure ethical or legal containment of harm arising from iterative, distributed, and adaptive technological processes. The result is not regulatory absence, but a regime in which legality and accountability increasingly diverge.

4. The Martens Clause: Ethical Corrective or Normative Illusion?

When technological developments outpace positive legal regulation, international lawyers frequently invoke the Martens Clause as a residual normative safeguard. First articulated in the 1899 Hague Convention II and reiterated in Additional Protocol I, the clause provides that in cases not covered by specific treaty provisions, civilians and combatants remain protected by "the principles of humanity and the dictates of the public conscience."[89] It is essentially a residual clause meant to fill in the gaps created by the passage of time and the evolution of humanity and its technology.

In contemporary debates on AWS, the Clause has re-emerged as a doctrinal refuge. Faced with uncertainty regarding the adequacy of Article 36 reviews and the limited applicability of export controls under the ATT, commentators and States alike cite the Martens Clause as evidence that autonomous weapons cannot escape humanitarian restraint merely because positive law is indeterminate.[90] Considering that the clause upholds principles of humanity and public conscience, it is argued to be capable of dealing with technological novelty and innovation in warfare. Even the ICJ, in its opinion, referred to the Martens Clause as an effective means of addressing the rapid evolution of military technology.[91] However, AI warfare seems far out of its reach.

---

[88] Vincent Boulanin, Kolja Brockmann and Luke Richards, *Responsible Artificial Intelligence Research and Innovation for International Peace and Security*, Stockholm International Peace Research Institute, November 2020, available at: https://www.sipri. org/sites/default/files/2020-11/sipri_report_responsible_artificial_intelligence_research_ and_innovation_for_international_peace_and_security_2011.pdf.

[89] Hague Convention II (1899) Preamble; Additional Protocol I (1977), art. 1(2); "Until a more complete code of the laws of war has been issued, the High Contracting Parties deem it expedient to declare that, in cases not included in the Regulations adopted by them, the inhabitants and the belligerents remain under the protection and the rule of the law of nations, as they result from the usages established among civilized peoples, from the laws of humanity and the dictates of public conscience".

[90] Rob Sparrow, "Ethics as a Source of Law: The Martens Clause and Autonomous Weapons", *Humanitarian Law & Policy*, 14 November 2017, available at: https://blogs.icrc. org/law-and-policy/2017/11/14/ethics-source-law-martens-clause-autonomous-weapons/.

[91] Advisory opinion, above note 31, p. 7.

Holistically, the clause is reactive rather than anticipatory, reflecting one influential yet highly contested interpretation of the Clause in contemporary IHL. Under this view, the Clause does not itself create new substantive prohibitions in advance of technological development; it operates as a residual interpretive safeguard once gaps in treaty law become apparent. However, this interpretation is not universally accepted. A persuasive argument is that Clause cannot, in itself, ensure state compliance in the context of autonomous weapon systems, which requires engaging with competing understandings of the Clause and explaining why the reactive reading is ultimately more convincing. Scholars have interpreted this provision in three ways: as a mere rhetorical statement, as an interpretive principle guiding the application of existing laws, or as an independent normative source capable of generating legal restraints beyond positive treaty law. Antonio Cassese's seminal article remains foundational because it maps these divergent interpretations and demonstrates that State practice has never conclusively settled the Clause's precise legal status.[92] Those advocating a proactive or anticipatory interpretation argue that the clause enables IHL to respond to emerging technologies before treaty law crystallizes. This position is especially visible in debates concerning AWS, where commentators and civil society actors invoke the Clause to argue that autonomous weapons remain constrained even in the absence of explicit treaty prohibition.[93]

Such approaches effectively treat the Clause as anticipatory because it purportedly enables legal restraint prior to formal legal codification. Yet this expansive reading faces serious doctrinal and practical difficulties. Firstly, State practice demonstrates little acceptance of the idea that the Martens Clause independently generates binding prohibitions. Cassese himself cautioned that although the Clause exerts moral and interpretive influence, States have generally resisted recognizing it as a self-sufficient source of concrete obligations. The International Court of Justice has likewise referred to the Clause primarily as an "effective means" of addressing humanitarian considerations rather than as an autonomous rule-creating mechanism, particularly in its Nuclear Weapons Advisory Opinion.[94] This suggests that the Clause supplements existing law rather than replacing the need for positive legal regulation. Secondly, the concepts of "humanity" and "public conscience" remain deeply indeterminate. In the AWS context, States sharply disagree over whether autonomous weapons inherently offend either standard. While some States advocate pre-emptive prohibition, others maintain that existing IHL is sufficient and continue investing heavily in

[92] Antonio Cassese, "The Martens Clause: Half a Loaf or Simply Pie in the Sky?", *European Journal of International Law,* Vol. 11(1), 187, 2000.
[93] Rebecca Crootof, "The Killer Robots Are Here: Legal and Policy Implications", *Cardozo Law Review*, Vol. 36(2), p. 1837, 2015.
[94] International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons,* Advisory Opinion*, ICJ Reports* 1996, 8 July 1996, para. 86.

autonomous military technologies.[95] Such disagreement undermines the claim that the Clause can function anticipatorily as a clear legal constraint. A norm whose content depends on contested moral evaluations is ill-suited to ensure consistent State compliance absent subsequent treaty development or consolidated customary practice. Accordingly, the stronger interpretation is that the Martens Clause operates reactively rather than anticipatorily. It becomes relevant once legal lacunae emerge and assists courts, tribunals, and States in interpreting humanitarian principles in situations insufficiently addressed by positive law.

There is nothing wrong with a clause being effective *ex post,* while the ATT is working *ex ante*. However, in the context of AI warfare, the gap the Martens clause was intended to address in the evolution of military technology is not adequately addressed by its content. Martens Clause cannot govern and or regulate AI warfare, it can only address the aftermath. Martens Clause, thus, cannot be over-relied upon to deal with modern military technology. The clause does not specify the level of accountability, how it must be enforced, or how the people will remain protected.[96] It is imperative to note that principles of humanity and public conscience remain undefined as well, perhaps intentionally. The more detrimental aspect is the lack of enforcement mechanisms. Martens clause works through interpretation, a guideline, within diplomatic discourse. No obligations are placed through the clause. A doctrinal link could be established under Article 36 read with the Clause.[97] It could be argued that this review under Article 36 could be informed by the Martens Clause. To explain, it means a review of new weapons should ensure that they are in line with the principles of humanity and public conscience.

5. From Compliance to Design: Reimagining State Responsibility through IHL-by-Design

IHL traditionally regulates weapons after they are designed, at the moment of use, and/or at the time of trade. The law explicitly regulates compliance and the aftermath with a presumption of concrete, tangible harm and concentrated responsibility. AI weapons topple the laws and regulations back on their heads. State responsibility gets diffused and dismantled with respect to AI warfare, under the existing IHL framework. The international order then breaks down: the powerful get more controlling, and the weak get weaker.

---

[95] ICRC, "ICRC Position on Autonomous Weapon Systems", 12 May 2022; and CCW Convention, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE on LAWS), "Draft articles on autonomous weapon systems—prohibitions and other regulatory measures on the basis of international humanitarian law (IHL)", Working paper submitted by Australia, Canada, Japan, the Republic of Korea, the United Kingdom and the United States, CCW/GGE/1/2023/WP.4, 6 Mar. 2023, p. 1.

[96] R. Sparrow, above note 90.

[97] Natalia Jevglevskaja, "Interpreting Article 36: The Object of Review" in *International Law and Weapons Review*, Cambridge University Press, Cambridge, 2021, p. 87.

In technology and AI governance discourse, "ethics-by-design" and "safety-by-design" are approaches that aim to build normative constraints into systems at the point of design, rather than relying solely on user behaviour, post-deployment regulation, or *ex post* liability.[98] IHL-by-design refers to embedding legal constraints into the architecture of weapons systems, rather than relying solely on *post hoc* compliance.[99] So, IHL-by-design implies a legal obligation of states, not a discretionary engineering choice.[100]

5.1. Re-engineering Article 36 and ATT Obligations

Article 36 imposes an *ex-ante* obligation, that applies during study, development, acquisition or adoption and requires review against all applicable international law, including IHL. It is apposite to note, that Article 36 does not limit itself to physical weapons and does not restrict the review to a one-time assessment. At least, not explicitly. ATT obligations address the trade in arms in a similar manner.

A very important reframing of the classification of weapons, inclusive of AI innovations, is to treat them as evolving systems. This is not just acceptable but defensible.[101] Article 36 traditionally assumes weapons are stable, predictable and static after deployment, rendering a single review sufficient.[102] AI systems break this assumption apart. AI-enabled weapons are update-driven, modular and adaptive. The re-engineering of Article 36 is proposed to be interpretive. The existing wording of Article 36 could be utilised for the re-engineering. The use of the words "study, development" could be interpreted to include retraining and upgrading of AI systems; "acquisition or adoption" to include integration into new platforms and "new weapon, means or method" to include evolved functional capabilities, not just objects. So, substantial software updates or capability expansions then retrigger review under Article 36.

There must be AI-specific metrics to address the risks associated with AI systems. Conventional weapons' risks assume that a weapon's harmful potential is materially fixed and that its effects can be assessed through relatively stable, observable characteristics. Risks associated with AI-systems, however, are systemic, opaque and probabilistic.[103] These characteristics directly affect core IHL principles, including distinction, proportionality, and precautions in attack, as codified in

---

[98] Lee Bygrave, "Security by Design: Aspirations and Realities in a Regulatory Context", *Oslo Law Review*, Vol. 8, No. 3, 2022, p. 126.
[99] R. Sparrow, above note 91, p. 24.
[100] Giovanna Maletta, "Seeking a Responsible Arms Trade to Reduce Human Suffering in Yemen", *Italian Journal of International Affairs*, Vol. 56, No. 1, 2021, p. 73.
[101] Matthis Maas, "Framing a Global AI Regime Complex in Five Steps", in *Architectures of Global AI Governance: From Technological Change to Human Choice*, Online Edition, Oxford Academic, Oxford, 2025, p. 497.
[102] V. Boulanin, above note 46, p. 12.
[103] Samuel Carey, "Regulating Uncertainty: Governing General-Purpose AI Models and Systemic Risk", *European Journal of Risk Regulation*, Vol. 1, 2025, p. 1.

Articles 48, 51, 52, and 57 of Additional Protocol I.[104] The ability to distinguish between civilians and combatants depends on reliable target identification; proportionality assessments require foreseeability of incidental harm; and precautions in attack demand the capacity to anticipate and mitigate risks. Where system behaviour is probabilistic, opaque, or context-dependent, traditional metrics are insufficient to evaluate compliance with these obligations.

While Article 36 addresses the regulation of weapons internally, the ATT addresses externalities. ATT deals with the supply and/or export of weapons. AI-enabled systems have software updates, cloud services, algorithms, data pipelines, etc. None of which qualify as arms, munitions or discrete export events. However, they still are lethal, have autonomy and enable repression.[105] The regulation of exports or trade, thus, must move beyond the tangible. It has to be capability-based. Articles 6 and 7, dealing with risk knowledge and the principle of overriding risk, should be used to review and re-assess whenever "capability" changes. Without extending export controls beyond the tangible, states or supplying entities get to evade responsibility. This shift allows for the accountability gap to be filled.

5.2. Implications for future AWS governance

The Convention on Certain Conventional Weapons (CCW)[106] regulates specific categories of conventional weapons considered to cause unnecessary suffering or indiscriminate harm. It has become the primary multilateral forum for discussions on LAWS.[107] The debates consider whether autonomous weapons should be banned, regulated, or addressed under existing IHL.[108]

The Hard Law has limitations regarding AWS governance.[109] AI systems undermine existing hard law because they are rapidly evolving, dual-use, and dependent on updates. Hard law, then, is obsolete, easily circumvented and outdated with respect to AI systems. Adaptive regulations, on the other hand, are not only compatible but ideal for AI systems. It refers to continuous review mechanisms, design-stage legal constraints, iterative compliance obligations, and post-deployment oversight. The suggestion to re-engineer Article 36 or to continue its ongoing review is an example of adaptive regulation. The impasse between the two highlights a

---

[104] Additional Protocol I, above note 11, p. 4.
[105] V. Boulanin, K. Brockmann and L. Richards, above note 88 p. 22.
[106] UNODA, "Convention on Certain Conventional Weapons", *UN-iLibrary*, March 2016, available at: https://www.un-ilibrary.org/content/books/9789210578967.
[107] Josephine Jackson, "Mapping the Lethal Autonomous Weapons Debate: An Introduction", *Ethics & International Affairs*, Vol. 37, No. 3, 2023, p. 254.
[108] Charukeshi Bhatt and Tejas Bharadwaj, "Understanding the Global Debate on Lethal Autonomous Weapons Systems: An Indian Perspective", *Carniege India*, 30 August 2014, available at: https://carnegieendowment.org/russia-eurasia/research/2024/08/understan ding-the-global-debate-on-lethal-autonomous-weapons-systems-an-indian-perspective.
[109] Damian Copeland, "How does International Law Regulate AWS?", in *A Functional Approach to the Legal Review of Autonomous Weapon System*, Brill, 2025, p. 41.

broader tension between hard law and adaptive regulation in the governance of autonomous weapons. While adaptive regulations seem ideal, strictly adaptive regulations would also risk fragmentation, uneven implementation, and erosion of legal accountability. Hard law offers clarity, legitimacy and enforceability. Thus, it is reliable and stable. AWS governance requires both flexibility and responsiveness, along with functional relevance, which are offered by adaptive regulations.

6. Concluding Remarks

IHL rests on the premise that decisions about life and death remain attributable to identifiable human or institutional actors who can be held responsible. If lethal decision-making is increasingly delegated to opaque or adaptive systems without corresponding mechanisms of control and review, this premise weakens. The analysis demonstrates that the problem confronting IHL is not the absence of regulation, but the misalignment between inherited legal architectures and contemporary technological realities.

Article 36 reviews should be reconceived as continuous and lifecycle-based. Reviews must extend beyond initial acquisition to encompass software updates, retraining processes, and operational modifications. Static certification is incompatible with adaptive technologies. This approach is already implicit in ICRC guidance and should be treated as a legal requirement rather than best practice. Article 36 presumes weapons whose behaviour can be assessed *ex ante*, AI-enabled systems evolve after deployment. The ATT regulates discrete transfers of tangible arms, and military AI diffuses through software, data, and distributed supply chains. The Martens Clause preserves humanitarian values but lacks operational precision. Each framework retains normative relevance, even in the absence of a binding treaty on AWS; shared interpretive principles and technical benchmarks can harmonise state practice and reduce regulatory arbitrage. Ultimately, the challenge posed by AI-enabled warfare is not whether machines can act lawfully, but whether states will accept responsibility for how those machines are built. The future credibility of international humanitarian law depends on that answer.

Human Dignity and Machine Agency: Reconciling International Humanitarian Law with the Ontology of Artificial Decision-Making

Sergey Sayapin[*]

ABSTRACT

As artificial intelligence (AI) increasingly mediates battlefield decision-making, international humanitarian law (IHL) faces a structural challenge – preserving its human-centred foundations in an algorithmic operational environment. This article introduces dignitary design as a jurisprudential framework for embedding respect for human dignity within the architecture, governance, and deployment of military AI systems. Grounded in the principle of humanity and the Martens Clause, dignitary design reconceptualises dignity as both a legal standard and a design requirement. The article argues that the legitimacy of future armed conflict regulation depends on integrating humanitarian values directly into technological systems, rather than relying solely on post hoc accountability.

Keywords: accountability – artificial intelligence (AI) – dignitary design – human dignity – international humanitarian law (IHL) – Martens Clause – military AI systems

I. Introduction: From Human Command to Algorithmic Mediation

Contemporary armed conflict is undergoing a profound structural transformation.[1] Artificial intelligence (AI) systems, once confined to auxiliary functions such as logistics optimisation, battlefield sensing, and intelligence analysis, are increasingly being integrated into operational processes that shape targeting, surveillance, force allocation, and strategic planning.[2] This development has generated growing concern among States, international organisations, military practitioners, and scholars regarding the future relationship between emerging technologies and the law of armed conflict.[3] As algorithmic systems become more deeply embedded in military decision-making, international humanitarian law (IHL) confronts a

[*]Dr Sergey Sayapin [s.sayapin@kimep.kz] is Professor of Law at KIMEP University (Almaty, Kazakhstan) and Distinguished Visiting Global Scholar at the NUS Centre for International Law (2025). The author thanks Dr. Samuel White and Dr. Jonathan Kwik for their helpful comments on an earlier draft of this article.

[1] ICRC, "International Humanitarian Law and the Challenges of Contemporary Armed Conflicts", *International Review of the Red Cross*, Vol. 106, No. 927, 2024.

[2] Wen Zhou and Anna Rosalie Greipl, "Artificial Intelligence in Military Decision-Making: Supporting Humans, not Replacing Them", *Humanitarian Law & Policy*, 29 August 2024, available at: https://blogs.icrc.org/law-and-policy/2024/08/29/artificial-intelligence-in-mil itary-decision-making-supporting-humans-not-replacing-them/ (all internet references accessed on 9 July 2026).

[3] ICRC and Geneva Academy, *Expert Consultation Report on AI and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts*, ICRC, Geneva, March 2024.

fundamental challenge: how can a legal regime built upon assumptions of human judgment, responsibility, and moral agency preserve its normative integrity when increasingly consequential decisions are mediated by computational systems?

Debates concerning military AI have expanded rapidly over the past decade. Much of the existing literature has focused on the regulation of lethal autonomous weapons systems (LAWS), the requirement of meaningful human control, [4] accountability for algorithmically mediated harms,[5] and the adequacy of existing weapons-review mechanisms under Article 36 of Additional Protocol I.[6] These discussions have made important contributions to understanding the legal and ethical implications of emerging technologies. Yet they have tended to focus on questions of compliance, attribution, and regulation, often treating AI as an external technological development to which existing legal frameworks must adapt. Comparatively less attention has been devoted to a more fundamental question: what happens to the human-centred normative structure of IHL when operational judgment itself becomes increasingly mediated by algorithmic systems?

Addressing this question requires greater conceptual precision regarding the technologies under discussion. Not every form of automation raises the concerns examined in this article. Military operations have long relied on technologies that mediate perception, communication, navigation, targeting, and damage assessment. Automated defensive systems, precision-guided munitions, and collateral damage estimation tools have been integral components of military decision-making for decades. The challenge posed by contemporary AI therefore does not arise from technological mediation as such – rather, it emerges when algorithmic systems begin to perform evaluative, predictive, and classificatory functions traditionally associated with human judgment.[7] Machine-learning systems capable of identifying patterns, generating recommendations, prioritising targets, predicting civilian presence, or shaping proportionality assessments do not merely execute pre-programmed instructions.[8] Instead, they increasingly participate in the production of operational knowledge upon which legally significant decisions are based. Thus, the legal significance of AI lies not in automation alone but in its growing role in structuring the informational and evaluative environment within which decisions concerning the use of force are made.

---

[4] Daniele Amoroso and Guglielmo Tamburrini, "Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues", *Current Robotics Reports*, Vol. 1, 2020.
[5] Gerald Mako, "Legal Accountability for AI-Driven Autonomous Weapons", *Articles of War*, 9 March 2026, available at: https://lieber.westpoint.edu/legal-accountability-ai-driven-autonomous-weapons/.
[6] Natalia Jeglevskaja, *International Law and Weapons Review: Emerging Military Technology under the Law of Armed Conflict*, Cambridge University Press, Cambridge, 2021, pp. 207–238.
[7] 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*, Vol. 107, No. 930, 2025.
[8] Natassha Selvaraj, "8 Machine Learning Models Explained in 20 Minutes", *Datacamp*, 16 September 2022, available at: https://www.datacamp.com/blog/machine-learning-models-explained.

This development creates a jurisprudential challenge that extends beyond familiar debates about compliance and accountability. At its core, IHL is not merely a technical regime regulating the conduct of hostilities – it is a normative system grounded in assumptions about human agency, moral responsibility, and the intrinsic value of human life. These commitments are reflected throughout the law's foundational principles, including distinction, proportionality, precaution, and humanity.[9] The prevailing legal discourse often assumes that existing IHL principles can simply be extended to AI-enabled systems through procedural safeguards, attribution rules, or human-supervision requirements.[10] While such measures are undoubtedly important, they may not fully address the deeper challenge posed by algorithmic mediation. Human decision-makers are increasingly required to rely upon outputs generated by systems whose internal reasoning processes may be difficult to understand, reconstruct, or meaningfully challenge. The concern is not that machines have become legal actors in their own right, nor that human decision-makers have disappeared from the battlefield[11] – rather, it is that algorithmic systems may progressively reshape how information is generated, interpreted, prioritised, and acted upon, thereby altering the conditions under which human judgment itself is exercised.

This article argues that preserving the legitimacy and vitality of IHL in an era of algorithmic warfare requires a reconceptualisation of the relationship between law, technology, and human dignity. Rather than treating dignity as a value protected only through *post hoc* accountability or judicial review, the article proposes that dignity should be understood as a design requirement shaping the architecture, governance, and deployment of military AI systems. This approach is conceptualised as "dignitary design". Dignitary design is not intended as a substitute for meaningful human control, nor does it replace existing obligations relating to weapons reviews, accountability, or compliance with the rules governing the conduct of hostilities. Instead, it provides a broader normative framework within which these mechanisms may be understood and developed. It proceeds from the premise that humanitarian values should not merely constrain the consequences of technological systems after deployment – they should be incorporated into the design, evaluation, governance, and operational use of those systems from the outset. In this sense, dignity becomes not only a legal value governing outcomes but also a structural principle shaping the conditions under which technological systems are developed and employed.

The urgency of this reframing arises from the growing gap between responsibility and agency in contemporary military operations. Traditional IHL

---

[9] For an overview of modern IHL, see Nils Melzer, *International Humanitarian Law: A Comprehensive Introduction*, ICRC, Geneva, 2019.

[10] Michael Pizzi, Mila Romanoff and Tim Engelhardt, "AI for Humanitarian Action: Human Rights and Ethics", *International Review of the Red Cross*, Vol. 102, No. 913, 2020.

[11] Aida Amanuel, "Removing the Human Decision-Making from Humanitarian Law An Examination of the Legal Implications of Autonomous Weapon Systems and Their Compliance with International Humanitarian Law", *Uppsala University*, Autumn Term 2024.

presumes that decisions to apply force are made by human actors capable of deliberation, contextual assessment, and moral evaluation, and legal accountability mechanisms are constructed accordingly: commanders assess proportionality, operators exercise precaution, legal advisers provide guidance, and States bear responsibility for internationally wrongful acts. AI complicates this architecture by distributing elements of cognition and assessment across complex socio-technical systems involving programmers, data curators, engineers, procurement authorities, military operators, commanders, and political institutions. As algorithmic systems increasingly contribute to operational judgments, responsibility becomes dispersed across multiple actors, while the practical locus of decision-making becomes more difficult to identify.[12]

This article does not advocate the prohibition of military AI, nor does it advance claims regarding machine consciousness or artificial legal personhood. Rather, it proceeds from a more modest but more demanding proposition: when States deploy systems capable of shaping decisions that may result in lethal effects, they bear an obligation to ensure that those systems are designed and governed in ways consistent with the normative commitments that underlie IHL. This requires moving beyond questions of compliance alone and considering how humanitarian values may be embedded within technological architectures themselves. From a jurisprudential perspective, such an approach represents an evolution rather than a departure from existing humanitarian law. Throughout its history, IHL has responded to technological change by reaffirming its underlying normative commitments while adapting their practical application to new means and methods of warfare. The principles of humanity and the dictates of public conscience, reflected most prominently in the Martens Clause,[13] provide a doctrinal foundation for extending humanitarian protections into domains not explicitly addressed by positive law. Artificial intelligence presents a particularly significant challenge because it affects not merely the means through which force is applied but the processes through which judgments concerning the use of force are formed. If IHL is to retain its human-centred identity under such conditions, dignity must be understood not only as a protected value but also as a principle capable of shaping technological design and governance.

The principal contribution of this article is therefore to reconceptualise human dignity as more than a moral aspiration or interpretive value within IHL. It argues that dignity should function as a design constraint that structures the development, deployment, and governance of military AI systems. In doing so, the article seeks to bridge existing debates concerning autonomy, accountability, and

---

[12] Sergey Sayapin, "International Law for a Fragile World: When Risk Outpaces Control – AI, Bioengineering, and Cyber-Autonomy", *Opinio Juris*, 23 April 2026, available at: https://opiniojuris.org/2026/04/23/international-law-for-a-fragile-world-when-risk-outpaces-control-ai-bioengineering-and-cyber-autonomy/.

[13] Antonio Cassese, "The Martens Clause: Half a Loaf or Simply Pie in the Sky", *European Journal of International Law*, Vol. 11, 2000.

human control while offering a framework capable of preserving the normative foundations of humanitarian law in an increasingly algorithmic battlespace.

The article develops this argument in seven sections. Section II examines the ontological status of artificial decision-making and its implications for legal agency. Section III situates human dignity within the normative structure of IHL, demonstrating its implicit centrality despite limited textual visibility. Section IV introduces the concept of dignitary design as a synthesis of law, ethics, and technology. Section V addresses accountability in hybrid human-machine systems and proposes expanded responsibility frameworks capable of addressing distributed agency. Section VI explores how dignity may be operationalised through proportionality assessments, precautionary obligations, weapons reviews, and algorithmic auditing. The conclusion reflects on the future legitimacy of IHL in a world where warfare is increasingly mediated by computational systems. Ultimately, the central question posed by military AI is not whether machines can comply with humanitarian law but whether humanitarian law can preserve its human-centred identity when operational judgment is increasingly shaped by machines. The answer depends upon whether dignity remains merely a rhetorical ideal or becomes a structural requirement of technological warfare.

II. The Ontological Problem: Can Machines "Decide" in Law?

The accelerating integration of artificial intelligence into military operations compels a foundational question that lies at the intersection of law, ethics, and technology: can machines meaningfully be said to "decide" in a juridical sense?[14] The answer bears directly on the integrity of IHL, which has traditionally assumed that legally significant acts in armed conflict originate in human deliberation and remain attributable to human actors. Without clarifying the ontological status of AI decision-making, contemporary debates concerning regulation, accountability, and compliance risk resting upon conceptual ambiguity.

At first glance, the language of "machine decision" appears metaphorical. Contemporary AI systems do not possess consciousness, intentionality, moral awareness, or the capacity for reflective judgment.[15] They process information through algorithmic architectures, generating outputs based on statistical inference, pattern recognition, and optimisation functions.[16] Even advanced machine-learning models do not currently deliberate in the manner associated with human practical

[14] Carlos Batallas, "When AI Meets the Laws of War", *Insights*, 3 October 2024, available at: https://www.ie.edu/insights/articles/when-ai-meets-the-laws-of-war/.
[15] Jeff Sebo and Robert Long, "Moral Considerations for AI Systems by 2030", *AI and Ethics*, Vol. 5, 2025.
[16] Antony Massobrio, "Machine Learning Optimization: Best Techniques and Algorithms", *Neural Concept*, 10 February 2023, available at: https://www.neuralconcept.com/post/ machine-learning-based-optimization-methods-use-cases-for-design-engineers.

reason.[17] Yet in operational settings, the outputs of such systems may shape target selection, collateral damage estimation, threat identification, force allocation, intelligence prioritisation, and other activities with profound legal consequences. When an algorithm classifies an object as a military objective, recommends engagement, or influences a proportionality assessment, the outcome is not metaphorical – it may contribute directly to decisions involving the use of lethal force.

At the same time, the challenge posed by AI should not be overstated. Military organisations have long relied upon technologies that mediate operational judgment.[18] Intelligence platforms, surveillance systems, predictive models, targeting software, and collateral damage estimation tools have influenced military decision-making for decades. The significance of contemporary AI therefore does not lie in the mere existence of technological mediation – rather, it lies in the increasing delegation of evaluative, predictive, and classificatory functions to systems capable of generating recommendations, prioritisation frameworks, and probabilistic assessments that shape human decision-making in ways that may be difficult to understand, contest, or reconstruct.[19] In other words, the novelty of AI resides less in automation as such than in the growing integration of machine-generated assessments into the informational architecture of military decision-making.

This distinction is important because not all forms of automation raise the same legal concerns. Conventional automated systems generally operate according to predefined rules and predictable parameters established by human designers.[20] By contrast, machine-learning systems may generate outputs through adaptive processes that are not fully transparent even to their developers.[21] The challenge posed by AI does not stem from the introduction of quantification into warfare. Quantitative reasoning has always formed part of military operations. Commanders routinely rely upon casualty estimates, intelligence assessments, risk matrices, operational models, and probabilistic forecasts. Proportionality analysis itself often incorporates predictive judgments regarding expected civilian harm and anticipated military advantage. The distinction between human and machine decision-making cannot

---

[17] J. M. Paterson, "AI Mimicking and Interpreting Humans: Legal and Ethical Reflections", *Journal of Bioethical Inquiry*, Vol. 22, No. 3, 2025.

[18] Patrick Hofstetter, Marius Geller and Florian Gerster, "The Impact of Artificial Intelligence On The Military Decision-Making Process and Mission Command", *The Defence Horizon Journal*, 11 September 2025, available at: https://tdhj.org/blog/post/ai-military-decision-making-2/.

[19] Tim Keating, Sarah Matar, Amr Fayad and Elena Chammas, "Preserving Human Moral Agency as use of AI-driven Autonomous Weapons Grows", *Kearney 100*, 15 May 2025, available at: https://www.kearney.com/industry/aerospace-defense/paris-air-show/artic le/preserving-human-moral-agency-as-use-of-ai-driven-autonomous-weapons-grows.

[20] Ajay Bandi, Bhavani Kongari, Roshini Naguru, Sahitya Pasnoor and Sri Vidya Vilipala, "The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges" *Future Internet*, Vol. 17, No. 9, 2025.

[21] Keonyoung Park and Ho Young Yoon, "AI Algorithm Transparency, Pipelines for trust not Prisms: Mitigating General Negative Attitudes and Enhancing Trust Toward AI", *Humanities and Social Sciences Communications*, Vol. 12, 2025.

therefore be reduced to a contrast between qualitative and quantitative reasoning. The more significant concern is whether algorithmic systems alter the relationship between calculation and judgment. When AI-generated outputs acquire an appearance of objectivity, precision, or authority, there is a risk that human decision-makers may defer to them without fully understanding their underlying assumptions, limitations, or uncertainties. The problem is not that machines perform calculations – rather, it is that machine-generated assessments may gradually become substitutes for independent human evaluation. Under such conditions, quantification risks evolving from a tool that informs judgment into a mechanism that structures or constrains it. The legal challenge therefore concerns not the existence of numerical reasoning but the possibility that computational outputs may increasingly shape the framework within which normative judgments are exercised.

One possible response is to insist that machines never truly decide and that responsibility therefore remains entirely with the human actors who deploy them.[22] Under this view, AI systems remain tools, however sophisticated, and all legally relevant decisions continue to originate with human operators, commanders, and States. This position possesses considerable intuitive appeal and remains broadly consistent with existing legal doctrine. Nevertheless, it risks becoming formalistic if applied without regard to the operational realities of contemporary military systems. In highly automated environments, human involvement may be temporally distant, informationally constrained, or structurally dependent upon algorithmic outputs. When commanders rely upon machine-generated assessments produced at speeds or levels of complexity beyond ordinary human cognitive capacities, their role may become increasingly supervisory rather than deliberative.[23]

A more accurate description of contemporary military AI is therefore neither replacement nor complete delegation, but hybridisation. Decision-making becomes distributed across complex socio-technical systems involving programmers, data curators, engineers, procurement authorities, legal advisers, commanders, operators, and algorithmic processes.[24] Programmers define architectures and optimisation objectives; data curators shape training datasets; engineers establish system parameters; legal advisers assess compliance; commanders determine operational goals; algorithms generate classifications and predictions; and operators execute or authorise actions. No single actor necessarily embodies the traditional image of autonomous judgment – instead, operational cognition becomes distributed across networks of human and technical actors. This phenomenon may be described as

[22] Robert Sparrow and Gene Flenady, "The Testimony Gap: Machines and Reasons", *Minds and Machines*, Vol. 35, 2025.
[23] Dmytro Shumlianskyi, "Some U.S. Military Personnel, Including Generals, Begin Using Generative Chatbots for Decision-Making", *Militarnyi*, 19 October 2025, available at: https://militarnyi.com/en/news/some-u-s-military-personnel-including-generals-begin-using-generative-chatbots-for-decision-making/.
[24] S. Sayapin, above note 12.

distributed agency.[25] The concept does not imply that machines possess legal personality or moral responsibility – rather, it recognises that legally significant outcomes increasingly emerge from interactions among multiple human and technical components. AI systems contribute to the production of operational knowledge but they do not become moral agents in their own right. Responsibility remains anchored in human and institutional actors even when the informational pathways leading to particular outcomes are shaped by algorithmic processes.[26]

This distinction is particularly important because some scholars have proposed forms of electronic personhood or artificial legal personality as responses to emerging technological challenges.[27] Such approaches do not resolve the concerns examined in this article. AI systems may improve situational awareness, reduce certain forms of human error, assist compliance with distinction obligations, and contribute to more informed operational decisions. The concern addressed here arises not from technological capacity itself but from the possibility that computational systems may reshape the conditions under which human judgment is exercised. The central question is therefore not whether machines can become moral agents – they cannot. Nor is it whether technology may assist compliance with IHL – it often can. The question is how legal systems can preserve meaningful human responsibility and moral judgment when operational cognition becomes increasingly distributed across human – machine assemblages.

This ontological challenge reveals the limitations of purely anthropocentric models of legal responsibility. Existing doctrines remain indispensable because responsibility must ultimately be borne by human and institutional actors. Yet those doctrines were developed in a context where legally significant decisions could generally be traced to identifiable individuals exercising relatively direct forms of judgment. AI complicates this picture by distributing cognition, knowledge production, and evaluative processes across complex socio-technical systems. The result is not the disappearance of human agency but its transformation. It is precisely this transformation that motivates the concept of dignitary design developed in the following sections. If artificial systems cannot internalise dignity as a moral value, dignity must instead be externalised as a normative constraint governing their design, deployment, and governance.[28] The task is therefore not to determine whether machines can become moral actors but to ensure that the technological systems through which military decisions are increasingly mediated remain structured by the humanitarian commitments that underlie IHL. The absence of machine morality thus becomes the reason for embedding human-centred safeguards into technological

---

[25] Werner Rammert, *Distributed Agency & Digital Technology: A Social Pragmatist View on Human – Technology Interaction*, Springer, Wiesbaden, 2025, pp. 129–142.

[26] Alexander Schwarz, "War Crimes", *Oxford Public International Law*, September 2014, available at: https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199 231690-e431.

[27] See, for example, Pawe Nowik, "Electronic personhood for artificial intelligence in the workplace", *Computer Law & Security Review*, Vol. 42, 2021.

[28] See section IV below.

architectures themselves. Dignitary design thus emerges not as a technological aspiration but as a normative response to the realities of distributed agency in contemporary warfare.

III. Human Dignity as a Structural Principle of International Humanitarian Law

Human dignity occupies an unusual position within IHL. Unlike international human rights law (IHRL), where dignity is frequently articulated as an explicit foundational value,[29] IHL more rarely invokes the concept directly.[30] This relative textual silence, however, should not be mistaken for normative absence. While dignity may not appear as a comprehensive organising principle in the treaty language of IHL in the same manner as it does in many human rights instruments, it nevertheless performs a structural function within the humanitarian legal order. It informs the interpretation of core principles, underlies protections afforded to persons affected by armed conflict, and provides part of the normative foundation upon which the legitimacy of lawful warfare depends. Understanding dignity in this manner is essential for assessing how humanitarian protections may be preserved under conditions of increasing algorithmic mediation.

The historical development of IHL reflects a persistent effort to reconcile military necessity with respect for the humanity of persons affected by armed conflict.[31] From the earliest Geneva instruments of the nineteenth century to the contemporary law of armed conflict, humanitarian regulation has sought not merely to limit violence but to preserve a minimum moral order within situations characterised by organised force. This aspiration is evident in the gradual expansion of protections for the wounded and sick, prisoners of war, civilians, and other protected persons. While these rules are often expressed in operational terms, they embody a deeper normative commitment – the recognition that even during armed conflict human beings retain an irreducible moral status that constrains the permissible exercise of violence.

In this respect, dignity occupies a position analogous to that which it enjoys in IHRL, albeit through different doctrinal pathways. Human rights law frequently presents dignity as the foundation from which specific rights derive.[32] Humanitarian law proceeds differently – rather than establishing dignity as an explicit source of rights, it operationalises respect for human worth through principles governing the

---

[29] Sergey Sayapin, "The International Committee of the Red Cross and International Human Rights Law", *Human Rights Law Review*, Vol. 9, No. 1, 18 February 2009, pp. 95–126; Sergey Sayapin, "International Committee of the Red Cross and the Use of International Human Rights Law", in D. Rogers (ed.), *Human Rights in War*, Springer, Singapore, 2022, pp. 217–23 6.

[30] See Article 3(1)(c) common to the 1949 Geneva Conventions.

[31] Michael N. Schmitt, "Military Necessity and Humanity in International Humanitarian Law: Preserving the Delicate Balance", *Virginia Journal of International Law*, Vol. 50, No. 4, 2010.

[32] Oscar Schachter, "Human Dignity as a Normative Concept", *American Journal of International Law*, Vol. 77, No. 4, 1983.

conduct of hostilities and the treatment of protected persons. Yet the underlying normative concern remains comparable – both legal regimes reject the proposition that human beings may be reduced entirely to instruments of political, military, or strategic objectives. This anti-instrumentalisation function is particularly important. At its most fundamental level, dignity affirms that individuals possess value beyond their utility to collective goals.[33] IHL does not prohibit violence as such but it seeks to ensure that persons affected by warfare are not treated merely as objects through which military advantage is pursued. This idea is reflected throughout the structure of IHL: civilians may not be targeted because they are civilians; prisoners of war must be treated humanely regardless of their military significance; persons *hors de combat* retain legal protection despite their loss of operational utility. These obligations are difficult to explain solely through considerations of efficiency, reciprocity,[34] or military necessity. They derive their normative force from the recognition that persons retain moral significance even under conditions of armed conflict.

The clearest doctrinal expression of this orientation is found in the principle of humanity. Humanity has long occupied a central position within the humanitarian tradition and remains one of the foundational principles of IHL. Although its precise contours remain contested, the principle broadly requires that warfare be conducted within limits that respect the human condition of those affected by conflict.[35] It prohibits the infliction of suffering unnecessary for the achievement of legitimate military objectives and serves as a normative constraint upon the exercise of force. Humanity is therefore more than an ethical aspiration – it performs an interpretive and justificatory function within IHL, helping to explain why certain forms of conduct are prohibited even when they may offer military advantage.

Closely connected to the principle of humanity is the Martens Clause, first formulated in the 9th preambular paragraph of the 1899 Hague Convention (II) with Respect to the Laws and Customs of War on Land and subsequently reaffirmed in numerous humanitarian instruments, including Article 1(2) of Additional Protocol I. The Clause provides that, in situations not specifically addressed by treaty law, civilians and combatants remain under the protection of the principles of international law derived from established custom, the principles of humanity, and the dictates of public conscience. Obviously, the Martens Clause does not provide a complete legal regime governing military AI, nor does it resolve all questions concerning autonomy, accountability, or algorithmic decision-making. What it does provide is doctrinal support for the proposition that humanitarian interpretation must remain responsive to emerging forms of warfare even when positive law has

[33] Remy Debes, "Dignity", *The Stanford Encyclopedia of Philosophy*, 18 February 2023, available at: https://plato.stanford.edu/archives/spr2023/entries/dignity.
[34] ICRC, International Humanitarian Law Databases, Rule 140: Principle of Reciprocity, available at: https://ihl-databases.icrc.org/en/customary-ihl/v1/rule140.
[35] ICRC, "What is IHL?", 18 September 2015, available at: https://www.icrc.org/en/ document/what-ihl.

not yet developed detailed regulatory frameworks. In this respect, the Clause reinforces the broader argument advanced in this article – that technological innovation does not displace the normative commitments of IHL but instead requires their application under new conditions.

The relationship between dignity and positive humanitarian law becomes particularly visible in the treatment of protected persons. Common Article 3 of the Geneva Conventions prohibits, among other conduct, outrages upon personal dignity, particularly humiliating and degrading treatment. Similar concerns are reflected in provisions governing prisoners of war, civilians under occupation, and persons deprived of liberty. These rules do more than prohibit physical harm – they recognise that individuals possess interests in respect, recognition, and humane treatment that survive the onset of armed conflict. The influence of dignity extends beyond the treatment of protected persons and into the conduct of hostilities itself. Although the principles of distinction, proportionality, and precaution are generally expressed in operational language, they presuppose underlying judgments concerning the value of human life and the limits of permissible violence. [36] Distinction reflects the proposition that civilians may not be treated merely as obstacles to military success. Proportionality assumes that civilian lives possess significance that cannot be disregarded in pursuit of military objectives. Precaution requires parties to take feasible measures to minimise civilian harm, reflecting a continuing obligation to account for the welfare of those exposed to the effects of military operations. These principles are often analysed in technical terms, yet their normative coherence depends upon an underlying commitment to the worth of human beings.

This point deserves particular emphasis because proportionality is sometimes understood primarily as a balancing exercise between military advantage and civilian harm. Such a description is accurate but incomplete. Proportionality is not simply a mechanism for optimising military outcomes – its legitimacy depends upon the premise that civilian lives possess intrinsic value and cannot be treated merely as variables within strategic calculations. Without this premise, proportionality risks collapsing into a purely utilitarian exercise in which human beings are reduced to components of military cost-benefit analysis. Dignity supplies part of the normative foundation that prevents such a collapse. It reminds decision-makers that civilian harm is not merely an operational variable but a matter of moral and legal significance.

The same observation applies to precaution. The obligation to take feasible precautions reflects more than a desire to improve operational efficiency – it embodies a continuing concern for the protection of persons who may be affected by military operations. Precaution therefore illustrates how dignity operates not only as a retrospective standard for evaluating conduct but also as a prospective principle

[36] See section VI below.

shaping decision-making before harm occurs. In this respect, dignity influences not only outcomes but also processes.

Yet the structural role of dignity becomes more difficult to sustain when decision-making is increasingly mediated by algorithmic systems. Human dignity resists complete quantification. The value of a civilian life cannot be fully captured through statistical models, confidence scores, risk matrices, or optimisation functions. While such tools may provide useful information, they do not exhaust the normative considerations that underlie humanitarian decision-making. The challenge posed by AI therefore does not arise because dignity suddenly becomes relevant – it arises because dignity is already embedded within the normative architecture of IHL, while the technologies increasingly used to support military decision-making operate primarily through forms of classification, prediction, and optimisation. This tension reveals the deeper significance of dignity for contemporary debates concerning military AI.[37] If dignity functions as a structural principle of IHL, then the challenge is not simply ensuring that algorithmic systems comply with specific legal rules. It is ensuring that the normative commitments those rules embody continue to influence decision-making processes under conditions of technological mediation. In other words, the question is no longer whether dignity remains relevant to humanitarian law.[38] It is how dignity can continue to shape operational judgment when that judgment is increasingly informed, structured, and sometimes constrained by computational systems.

The answer cannot lie solely in *post hoc* accountability. Once dignity is recognised as a structural principle of IHL, it becomes necessary to consider how respect for human worth may be incorporated into the architecture of military technologies themselves. The challenge is therefore not merely interpretive but institutional and technological. This insight provides the foundation for the concept of dignitary design developed in the following section. If dignity already forms part of the normative architecture of humanitarian law, then preserving the human-centred character of IHL in an era of algorithmic warfare requires ensuring that dignity is reflected not only in legal evaluation after decisions are made but also in the systems through which those decisions are increasingly mediated.[39]

---

[37] Brian Stauffer, "A Hazard to Human Rights: Autonomous Weapons Systems and Digital Decision-Making", *Human Rights Watch*, 28 April 2025, available at: https://www.hrw.org/report/2025/04/28/a-hazard-to-human-rights/autonomous-weapons-systems-and-digital-decision-making.

[38] Ginevra Le Moli, "AI vs Human Dignity: When Human Underperformance is Legally Required", *Groupe d'études géopolitiques*, available at: https://geopolitique.eu/en/ articles/ai-vs-human-dignity-when-human-underperformance-is-legally-required/.

[39] ICRC, "Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach", *International Review of the Red Cross*, Vol. 102, No. 913, 2020, pp. 463–479.

IV. From Norm to Architecture: The Concept of Dignitary Design

The preceding sections have argued that military AI presents a challenge not because machines become moral agents but because algorithmic systems increasingly mediate the conditions under which human judgment is exercised. They have also argued that human dignity functions as a structural principle of IHL, informing the interpretation and application of core humanitarian norms even where it is not expressly articulated. The question that follows is therefore practical as well as jurisprudential: how can the normative commitments embodied in dignity continue to shape military decision-making when operational cognition is increasingly distributed across human – machine assemblages?

Existing legal and policy debates offer several partial answers: meaningful human control seeks to preserve human involvement in decisions concerning the use of force;[40] Article 36 weapons reviews provide mechanisms for assessing the legality of new weapons, means, and methods of warfare;[41] responsible AI frameworks emphasise transparency, explainability, fairness, and accountability; [42] and institutional accountability mechanisms seek to ensure that legal responsibility remains attributable to human actors and States. [43] Each of these approaches addresses an important dimension of the problem – yet none of them fully explains why preserving human-centred decision-making matters from the perspective of humanitarian law:

- meaningful human control, for example, identifies a procedural safeguard but does not itself provide a comprehensive normative account of what that safeguard is intended to protect;
- weapons reviews evaluate legality but do not necessarily address the broader humanitarian considerations embedded within technological systems;
- accountability frameworks determine who may be responsible when harm occurs but do not always guide how systems should be designed before deployment;
- responsible AI principles often emerge from broader governance debates and may not be tailored to the distinctive normative structure of IHL. As a result, existing approaches frequently operate

[40] Masoud Zamani, "How Meaningful is 'Meaningful Human Control' in LAWS Regulation?", *Articles of War*, 26 March 2025, available at: https://lieber.westpoint.edu/ how-meaningful-is-meaningful-human-control-laws-regulation/.

[41] *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*, ICRC, 2006, available at: https://library.icrc.org/library/docs/DOC/icrc_002_0902.pdf.

[42] Emmanouil Papagiannidis, Patrick Mikalef and Kieran Conboy, "Responsible Artificial Intelligence Governance: A Review and Research Framework", *The Journal of Strategic Information Systems*, Vol. 34, No. 2, 2025.

[43] Claudio Novelli, Mariarosaria Taddeo and Luciano Floridi, "Accountability in Artificial Intelligence: What it is and how it works", *AI & Society*, Vol. 39, 2023.

> as discrete mechanisms rather than as components of a unified humanitarian framework.

This article proposes dignitary design as such a framework. Dignitary design may be defined as the integration of respect for human dignity into the architecture, governance, evaluation, and operational deployment of military AI systems, with the objective of ensuring that technological mediation does not undermine the human-centred normative commitments of IHL. It seeks to preserve the role of dignity not merely as a value protected through legal review after harm has occurred but as a principle shaping the conditions under which military decisions are generated, informed, and implemented. Importantly, dignitary design does not require machines to recognise, understand, or internalise dignity – AI systems lack the moral capacities necessary for such recognition. They do not possess conscience, empathy, or the ability to appreciate the intrinsic worth of human beings. The objective of dignitary design is therefore not to create morally aware machines – rather, it is to ensure that human institutions embed dignity-sensitive constraints within the technological systems they develop and deploy. Precisely because machines cannot understand dignity, responsibility for preserving dignity must remain with the human actors and institutions that design, govern, and employ those systems.

Dignitary design thus differs from approaches that focus exclusively on outcomes. Traditional accountability frameworks often evaluate whether particular harms can be attributed to identifiable actors after an event has occurred, whereas dignitary design shifts attention upstream. It asks how technological systems should be structured before deployment, in order to preserve the humanitarian values that IHL seeks to protect. In this respect, it resembles the broader movement toward values-based design in technology governance, [44] while remaining grounded specifically in the normative commitments of humanitarian law. At its core, dignitary design rests upon a simple proposition: if dignity constitutes a structural principle of IHL, then military AI systems should be designed, governed, and deployed in ways that reflect that principle. The practical implications of this proposition may be organised into three interrelated dimensions – architectural, procedural, and institutional.

A. Architectural Dimension

The architectural dimension concerns the design characteristics of AI systems themselves. It focuses on how technological systems generate, present, and structure information relevant to military decision-making. If dignity requires that persons not be reduced to mere objects of optimisation, then the architecture of military AI systems should avoid treating civilian harm as a purely technical variable detached from its humanitarian significance. One implication is the importance of explainability. Human operators must be able to understand, at least to a meaningful

[44] See Gjalt de Graaf, *Value-Based Governance*, Cambridge University Press, Cambridge, 2025.

degree, the basis upon which significant recommendations or classifications are generated.[45] Explainability does not require complete transparency in every technical respect, nor does it imply that complex machine-learning models can always be rendered fully interpretable. It does, however, require that decision-makers possess sufficient information to evaluate, challenge, and contextualise algorithmic outputs. Systems that generate recommendations without providing meaningful opportunities for scrutiny risk transforming human oversight into a procedural formality.

The architectural dimension also encompasses the management of uncertainty.[46] Military AI systems frequently operate in environments characterised by incomplete information, rapidly changing conditions, and ambiguous data. Dignitary design therefore requires that uncertainty be made visible rather than concealed – that is, confidence scores, probability estimates, and margins of error should be presented in ways that encourage critical evaluation rather than passive acceptance. Human operators should be reminded not only of what a system predicts but also of the limitations of those predictions.

Similarly, system design should incorporate mechanisms for identifying and mitigating bias. [47] Training datasets, classification models, and optimisation objectives may systematically distort operational assessments in ways that increase the risk of civilian harm. Because dignity requires that individuals not be treated as expendable variables within computational processes, system architectures should be subject to ongoing testing, validation, and auditing designed to identify discriminatory or systematically harmful patterns.

Finally, architectural design should preserve meaningful opportunities for human intervention.[48] Override mechanisms, escalation procedures, and safeguards against fully automated implementation of high-consequence recommendations all serve to reinforce the continuing role of human judgment. The objective here is not to prevent technological assistance but to ensure that human actors retain genuine capacity to evaluate and, where necessary, reject algorithmic outputs.

---

[45] Markus Christen, Thomas Burri, Serhiy Kandul and Pascal Vörös, "Who is Controlling Whom? Reframing "Meaningful Human Control" of AI systems in Security", *Ethics and Information Technology*, Vol. 25, 2023.

[46] Jonathan Kwik, "Iterative Assessment for Military Artificial Intelligence (AI) Systems", in Bérénice Boutin, Taylor Kate Woodcock and Sadjad Soltanzadeh (eds), *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*, TMC Asser Press / Springer, 2026.

[47] Laura Bruun and Marta Bo, *Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law*, Stockholm International Peace Research Institute, August 2025.

[48] "The Risks and Opportunities of AI on Humanitarian Action", *Wilton Park*, 15–17 May 2024, available at: https://www.wiltonpark.org.uk/reports/the-risks-and-opportunities-of-ai-on-humanitarian-action-report/.

B. Procedural Dimension

The procedural dimension concerns the processes through which military AI systems are evaluated, reviewed, tested, and monitored throughout their operational lifecycle. Existing legal obligations already provide important foundations for such processes. Most notably, Article 36 of Additional Protocol I requires States to determine whether the employment of new weapons, means, or methods of warfare would be prohibited under international law. Yet the emergence of adaptive and data-driven systems raises questions that extend beyond traditional weapons reviews. Unlike many conventional weapons, AI-enabled systems may evolve over time through software updates, retraining processes, changing datasets, and new operational environments.[49] Their behaviour may therefore depend not only upon their original design but also upon subsequent interactions with data and users. Dignitary design responds to this reality by emphasising lifecycle governance rather than one-time review.

Under a dignitary design framework, legal review becomes an ongoing process rather than a singular event. Systems should be subject to continuous assessment before, during, and after deployment. Testing should examine not only technical performance but also the ways in which algorithmic outputs influence human decision-making. Auditing mechanisms should evaluate whether systems systematically encourage excessive reliance upon automated recommendations or diminish opportunities for meaningful human evaluation.[50] In this respect, dignitary design extends rather than replaces existing review mechanisms. Article 36 reviews, legal assessments, operational testing, and responsible AI evaluations remain indispensable. The contribution of dignitary design lies in providing a normative lens through which these processes may be understood and integrated.

C. Institutional Dimension

The institutional dimension concerns the allocation of responsibilities among the actors who shape military AI systems. The concept of distributed agency developed in section II suggests that legally significant outcomes increasingly emerge from interactions among multiple human and technical components. Preserving dignity under such conditions requires institutional arrangements capable of addressing this complexity.

Responsibility cannot rest exclusively with commanders or operators: developers influence system capabilities through design choices; data curators shape the informational foundations upon which systems learn and operate; procurement authorities determine which technologies enter operational environments; legal

---

[49] "AI Model Retraining", *Transforming Data with Intelligence (TDWI)*, available at: https://tdwi.org/pages/glossary/ai-model-retraining.aspx.

[50] "Independent AI Audit Requirements", *VerifyWise*, available at: https://verifywise.ai/lexicon/independent-ai-audit-requirements.

advisers assess compliance and identify risks; and political authorities establish strategic priorities and governance frameworks. Each of these actors contributes to the conditions under which algorithmic systems generate operational consequences. Legal advisers occupy a notable position within this framework.[51] Institutional responsibility also requires transparency and oversight mechanisms capable of generating organisational learning. Post-strike assessments,[52] incident investigations, and review procedures should not focus exclusively on assigning blame after failures occur. They should also identify structural weaknesses within socio-technical systems and facilitate improvements designed to reduce future risks.

D. Dignity as a Design Constraint

Taken together, these architectural, procedural, and institutional dimensions reveal the distinctive contribution of dignitary design. Meaningful human control, Article 36 review, explainability, auditing, and accountability remain indispensable mechanisms. Dignitary design does not replace them – rather, it provides the normative framework that explains why they matter and how they should be oriented. The central insight is that dignity should function as a design constraint rather than merely a retrospective evaluative standard. Traditionally, dignity enters legal analysis after harm has occurred, informing assessments of responsibility, legality, or reparation. Dignitary design shifts attention to an earlier stage – it asks whether technological systems are structured in ways that preserve the status of human beings as subjects of legal concern rather than objects of computational optimisation. This shift has significant implications for the future development of military AI. It suggests that compliance with IHL cannot be evaluated solely by examining outcomes – the processes through which decisions are generated also matter. A system that consistently obscures uncertainty, discourages scrutiny, or transforms civilian lives into abstract variables may undermine humanitarian values even if particular outcomes remain formally lawful. Conversely, systems designed to facilitate critical evaluation, preserve meaningful human judgment, and foreground humanitarian considerations may strengthen the capacity of military institutions to comply with IHL under increasingly complex conditions.

Dignitary design therefore represents neither a technological solution nor a new legal doctrine. It is a jurisprudential framework that seeks to align technological development with the normative commitments that have long animated IHL. By embedding dignity within the architecture, governance, and deployment of military AI systems, it offers a means of preserving the human-centred identity of IHL in an era of algorithmic warfare.

---

[51] See section V below.

[52] Jan Frelin, *Purposeful Assessment: Assessing the Effects of Military Operations*, FOI, Swedish Defence Research Agency, Stockholm, January 2009.

V. Accountability in Hybrid Human – Machine Assemblages

The emergence of military AI is frequently presented as an accountability problem. As algorithmic systems assume increasingly significant roles in targeting, surveillance, intelligence analysis, and operational planning, concerns arise regarding who should bear responsibility when harmful outcomes occur.[53] These concerns are understandable – IHL has traditionally been premised upon the attribution of legally significant acts to identifiable human actors and States. Yet contemporary military AI complicates this picture by distributing elements of cognition, assessment, and decision-support across complex socio-technical systems. The challenge is therefore not merely one of legal attribution – it is a question of how responsibility should be structured in operational environments characterised by distributed agency. The accountability challenge therefore lies not in identifying a single replacement for the traditional decision-maker but in understanding how responsibility should be allocated within increasingly complex networks of human and technical actors.

This perspective requires moving beyond simplistic narratives suggesting that AI creates accountability vacuums. The notion of an accountability gap has become a recurring theme within debates concerning autonomous systems.[54] According to this view, harmful outcomes may become difficult to attribute because no individual actor exercises complete control over the processes leading to a particular decision. While such concerns should not be dismissed, the language of accountability gaps may be misleading. Military operations have rarely been the product of isolated individual choices – even before the emergence of contemporary AI, decisions concerning the use of force depended upon multiple actors, institutions, information sources, and chains of authority. What AI does is not create complexity from nothing – rather, it renders existing complexity more visible while introducing additional layers of technological mediation. This observation suggests that accountability in military AI should be understood not primarily as a problem of absence but as a problem of distribution. In this respect, military AI may be understood through the lens of the "many-hands problem" – a concept developed in fields such as technology ethics, public administration, and risk governance.[55] The many-hands problem arises when harm results from the actions of multiple actors, none of whom individually controls the entire process leading to a particular outcome. Under such conditions, responsibility may become fragmented, diluted, or obscured even though numerous actors contribute to the relevant chain of events.

---

[53] Nuno Marques, "Artificial Intelligence, Causation and State Responsibility in Cyberspace", *International Law Blog*, 23 March 2026, available at: https://internationallaw.blog/2026/03/23/artificial-intelligence-causation-and-state-responsibility-in-cyberspace/.

[54] Peter Königs, "Artificial Intelligence and Responsibility Gaps: What is the Problem?", *Ethics and Information Technology*, Vol. 24, 2022.

[55] Matthijs M. Maas, *Architectures of Global AI Governance: From Technological Change to Human Choice*, Oxford University Press, Oxford, 2025, pp. 497–593.

Military AI presents a particularly clear example of this phenomenon: developers influence system behaviour through design choices and optimisation objectives;[56] data curators shape training datasets that affect classification accuracy and predictive performance; procurement authorities determine which technologies are acquired and deployed; legal advisers assess compliance with humanitarian law; commanders establish operational objectives and rules of engagement;[57] operators decide whether and how to act upon algorithmic recommendations; and political authorities define broader strategic priorities. Each actor contributes to the operational environment within which AI systems generate legally significant effects – yet no single actor necessarily exercises complete control over the entire process. Recognising the distributed nature of responsibility does not require abandoning traditional accountability frameworks. On the contrary, existing doctrines remain indispensable – individual criminal responsibility,[58] command responsibility,[59] and State responsibility continue to provide essential mechanisms for enforcing compliance with IHL.[60] However, these doctrines were largely developed in contexts where legally significant conduct could be more readily linked to identifiable human decisions.[61] AI does not render these frameworks obsolete but it does expose some of their limitations when applied to highly distributed decision-making environments.

The limitations of traditional accountability models become particularly apparent when responsibility is conceptualised exclusively in terms of individual fault. Consider a military AI system that incorrectly classifies a civilian structure as a military objective.[62] The resulting harm may reflect a combination of factors – flaws in training data, design assumptions, operational pressures, inadequate oversight, inappropriate deployment conditions, or excessive reliance upon algorithmic recommendations. Focusing exclusively upon the final operator risks obscuring the broader institutional and technological context within which the error occurred. Conversely, dispersing responsibility too broadly risks creating a situation in which no actor is meaningfully accountable. The challenge is therefore to develop

---

[56] Hayden Horner, "Engineers Navigating Ethics of AI Development in Line with UNESCO Principles", *Engineering Institute of Technology*, 3 June 2025, available at: https://www.eit.edu.au/engineers-ethics-ai-development-unesco-principles/.

[57] Ann-Katrien Oimann and Adriana Salatino, "Command Responsibility in Military AI Contexts: Balancing Theory and Practicality", *AI Ethics*, Vol. 5, 2025.

[58] See Ciara Damgaard, *Individual Criminal Responsibility for Core International Crimes: Selected Pertinent Issues*, Springer, Berlin / Heidelberg, 2008.

[59] Yoram Dinstein, "Command Responsibility", *Oxford Public International Law*, September 2015, available at: https://opil.ouplaw.com/display/10.1093/law:epil/97801992 31690/law-9780199231690-e273.

[60] Sergey Sayapin, *A Central Asian Perspective on International Law*, Hart Publishing, 2026, p. 219.

[61] Sergey Sayapin, "International Law for a Fragile World: Global Risk and the Limits of Classical International Law", *Opinio Juris,* available at: https://opiniojuris.org/2026/ 02/06/international-law-for-a-fragile-world-global-risk-and-the-limits-of-classical-international-law/.

[62] Jimena Sofía Viveros Álvarez, "The Risks and Inefficacies of AI systems in Military Targeting Support", *Humanitarian Law & Policy*, 4 September 2024, available at: https:// blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/.

accountability structures capable of recognising distributed causation without allowing responsibility to dissipate. This challenge is particularly significant because AI systems lack the moral capacities necessary for legal responsibility. Machines do not possess intentions, understand legal obligations, or appreciate the consequences of their actions in the manner required for criminal or moral accountability – consequently, proposals to attribute responsibility directly to AI systems provide little assistance in addressing humanitarian concerns. The problem confronting IHL is not the absence of an appropriate legal category for machines – it is ensuring that responsibility remains meaningfully connected to the human and institutional actors who design, govern, and deploy technological systems.

For this reason, State responsibility assumes heightened importance in the context of military AI. Under international law, States remain responsible for internationally wrongful acts attributable to them, regardless of whether those acts involve advanced technologies or traditional means of warfare.[63] The deployment of AI systems does not diminish these obligations. If anything, it reinforces the need for robust institutional safeguards capable of ensuring compliance with humanitarian norms. States cannot avoid responsibility by pointing to algorithmic complexity, software errors, or technological opacity – the use of AI therefore strengthens rather than weakens the importance of institutional due diligence.[64] The latter may be understood as the obligation to establish governance structures capable of identifying, managing, and mitigating risks associated with military AI. Such obligations extend beyond the narrow question of whether a particular system is technically functional – they encompass procurement procedures, testing protocols, legal review mechanisms, auditing processes, training requirements, operational oversight, and post-deployment evaluation. From the perspective of dignitary design, due diligence serves as a means of ensuring that humanitarian considerations remain embedded throughout the lifecycle of military AI systems.

This broader conception of responsibility also highlights the importance of legal advisers.[65] Article 82 of Additional Protocol I requires States to ensure that legal advisers are available to advise military commanders on the application of IHL. Traditionally, this function has focused on operational compliance and legal review. In the context of military AI, however, legal advisers assume a more expansive role. Their expertise may contribute not only to the assessment of particular military operations but also to the development of governance frameworks governing technological systems themselves. They may participate in weapons reviews, contribute to procurement decisions, evaluate testing procedures, advise on operational safeguards, assist in developing auditing mechanisms, and help identify risks associated with algorithmic decision-making. Their role is therefore not limited

[63] See ARSIWA, Article 1.
[64] Beyza Unal and Ulysse Richard, *Governance of Artificial Intelligence in the Military Domain*, UNODA Occasional Papers, No. 42, United Nations, New York, June 2024, pp. vi, 23.
[65] Leslie C. Green, "The Role of Legal Advisers in the Armed Forces", *Israel Yearbook on Human Rights*, Vol. 7, 1977.

to determining whether a system satisfies minimum legal requirements – it includes helping institutions understand how technological choices may affect the human-centred values embedded within IHL.

Notably, the importance of institutional governance extends beyond legal advisers alone – effective accountability requires organisational structures capable of generating transparency, oversight, and learning.[66] Dignitary design provides a useful framework for understanding these responsibilities because it links accountability to the preservation of human dignity rather than merely to the allocation of blame. Institutions are therefore responsible not merely for preventing unlawful outcomes but for ensuring that technological systems are governed in ways consistent with the humanitarian values that underlie IHL.

Ultimately, the challenge posed by military AI is not that responsibility disappears – it is that responsibility becomes more difficult to organise and more important to maintain. The preservation of dignity under conditions of distributed agency therefore depends upon governance structures capable of ensuring that responsibility remains visible throughout the lifecycle of military AI systems. If section IV demonstrated that dignity should function as a design constraint, the next question becomes how that constraint may be operationalised within specific decision-making processes. It is to that question – particularly in relation to proportionality, precaution, auditing, and operational review – that the article now turns.

VI. Operationalising Dignity: Proportionality, Precaution, and Algorithmic War

If dignity is to function as a meaningful constraint on military AI rather than a purely aspirational value, it must be translated into operational practices capable of guiding decision-making in concrete contexts. The challenge is not simply one of legal interpretation – it is a question of institutional design, operational governance, and technological architecture. The preceding sections have argued that dignity functions as a structural principle of IHL and that military AI increasingly mediates the conditions under which humanitarian judgments are exercised. The practical question that follows is therefore how dignity may continue to influence decision-making processes when those processes increasingly incorporate algorithmic systems.

A. Dignity and Proportionality in Algorithmic Warfare

The principle of proportionality provides a useful illustration of the challenges posed by AI-mediated decision-making. Proportionality prohibits attacks expected to cause incidental civilian harm that would be excessive in relation to the concrete and direct

[66] Daniel E. Esser and Heiner Janus, "How are Accountability and Organisational Learning Related? A Study of German Bilateral Development Cooperation", *Evaluation*, Vol. 29, No. 4, 2023.

military advantage anticipated.[67] The application of this principle necessarily requires predictive judgments under conditions of uncertainty. Commanders routinely assess intelligence, estimate potential civilian casualties, evaluate operational alternatives, and weigh anticipated military benefits – quantification has therefore always formed part of proportionality analysis. Contemporary military decision-making already relies upon casualty estimates, collateral damage estimation tools, risk assessments, and probabilistic judgments.[68] The challenge posed by AI lies elsewhere – it arises when algorithmic outputs begin to shape proportionality assessments in ways that obscure uncertainty, discourage independent evaluation, or transform predictive calculations into *de facto* decision rules. AI systems may contribute valuable information to proportionality assessments by identifying patterns, modelling potential effects, or processing large volumes of operational data.[69] Yet proportionality remains fundamentally a normative judgment – no algorithm can ever determine whether anticipated civilian harm is legally "excessive" in relation to expected military advantage without relying upon assumptions, proxies, and evaluative choices embedded within its design. The risk is therefore not that AI introduces quantification into proportionality analysis but that quantitative outputs may acquire an appearance of objectivity that discourages critical scrutiny.

From the perspective of dignitary design, proportionality assessments should preserve meaningful opportunities for human judgment. Human decision-makers should be able to understand how relevant outputs were generated, identify sources of uncertainty, evaluate alternative courses of action, and challenge algorithmic recommendations where appropriate. Dignity is thus protected not by excluding quantitative tools from military planning but by ensuring that quantitative assessments remain embedded within a broader framework of humanitarian evaluation. This concern is particularly important because dignity functions, in part, as a safeguard against the reduction of human beings to variables within optimisation processes – civilian lives cannot be understood solely as data points within a cost-benefit calculation. While proportionality necessarily involves balancing, the legitimacy of that balancing exercise depends on recognising the intrinsic value of those who may be harmed – dignitary design requires that military AI systems avoid presenting civilian harm as a purely technical parameter detached from its humanitarian significance.

---

[67] ICRC, "The Principle of Proportionality", available at: https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/04_proportionality-0.pdf.

[68] Zena Assaad, "Data, Decisions, and Death: Examining the Role of AI in Contemporary Warfare", *Arab Reform Initiative*, 27 May 2026, available at: https://www.arab-reform .net/publication/data-decisions-and-death-examining-ai-in-contemporary-warfare/.

[69] "Key Takeaways of The Military AI, Peace & Security Dialogues 2025", *UNODA*, 17 September 2025, available at: https://disarmament.unoda.org/en/updates/key-takeaways-military-ai-peace-security-dialogues-2025.

B. Precaution and the Governance of Technological Risk

The principle of precaution provides a second important avenue through which dignity may be operationalised.[70] Unlike proportionality, which focuses on balancing competing considerations in particular operational contexts, precaution is inherently future-oriented. It requires parties to take feasible measures designed to minimise civilian harm before attacks occur. In this respect, precaution is especially relevant to the governance of emerging technologies.

Military AI introduces new forms of uncertainty that may not be fully captured by traditional operational planning – as discussed above, AI performance may vary across environments, datasets may contain hidden biases, system behaviour may change following updates, and operators may develop inappropriate reliance upon algorithmic recommendations. These risks cannot always be addressed through retrospective accountability mechanisms, and require anticipatory governance – dignitary design therefore encourages a precautionary approach to military AI. In particular, it requires institutions to identify foreseeable risks and implement measures designed to reduce them – testing, verification, validation, simulation, red-teaming,[71] scenario-based evaluation, and independent review all serve this function. These practices help ensure that systems are assessed not only for technical reliability but also for their potential effects on humanitarian decision-making. Significantly, the precautionary dimension of dignitary design also extends to operational deployment – restrictions on the use of particular systems, limitations on operational environments, requirements for human review, and mechanisms for suspending deployment where performance becomes unreliable may all constitute precautionary measures. Such safeguards recognise that human dignity may be threatened not only by unlawful outcomes but also by institutional failures to anticipate foreseeable risks.

C. Dignity-Based Auditing

One of the most practical implications of dignitary design is the development of dignity-based auditing mechanisms. Traditional evaluations of military technologies often focus on effectiveness, reliability, accuracy, and legality.[72] While these considerations remain essential, they may not fully capture the broader humanitarian implications of AI-mediated decision-making. A dignity-based audit should ask not only whether a system functions as intended but also, importantly, whether it

---

[70] ICRC, International Humanitarian Law Databases, Rule 15: Principle of Precautions in Attack, available at: https://ihl-databases.icrc.org/en/customary-ihl/v1/rule15.

[71] "What is Red Teaming? Methodology and Scope of a Red Team Operation", *Vaadata*, 7 April 2026, available at: https://www.vaadata.com/en/blog/what-is-red-teaming-method ology-and-scope-of-a-red-team-operation/.

[72] Netta Goussac and Rain Liivoja, "Legal Reviews of Military Artificial Intelligence Capabilities", *Asia-Pacific Institute for Law and Security (APILS)*, 25 August 2025, available at: https://apils.org/2025/08/25/legal-reviews-of-military-artificial-intelligence-capabilities/.

preserves the conditions necessary for meaningful human judgment. Such an audit might examine, in particular, whether uncertainty is communicated transparently, whether algorithmic outputs can be challenged and reviewed, whether civilian protection is adequately represented within system design, and whether operators retain genuine opportunities to exercise independent evaluation. Similarly, dignity-based auditing should assess the extent to which systems incorporate safeguards against bias, overreliance, and inappropriate automation. It should also examine whether confidence scores are presented responsibly, whether alternative interpretations of data remain visible, and whether system architectures encourage critical engagement rather than passive acceptance of algorithmic recommendations.

D. Rules of Engagement and Operational Governance

Dignitary design also has implications for operational governance. Rules of engagement (ROE) have long served as mechanisms through which legal obligations are translated into operational guidance.[73] As AI systems become more deeply integrated into military decision-making, existing ROE frameworks may require adaptation to address new forms of technological mediation. AI-specific ROE might establish conditions governing when algorithmic recommendations may be relied upon, when additional human review is required, and when deployment should be suspended. Their rules could specify confidence thresholds below which recommendations cannot be acted upon without further verification. They might require escalation procedures where uncertainty exceeds predefined limits or where civilian harm cannot be assessed with sufficient confidence. Operational governance may also require restrictions on the contexts in which particular systems may be used – certain environments may present levels of complexity, uncertainty, or civilian presence that render heavy reliance on AI especially problematic. Dignitary design therefore supports context-sensitive governance rather than uniform assumptions regarding the suitability of particular technologies. More broadly, operational governance should preserve opportunities for human intervention throughout the decision-making process.

E. From Weapons Review to Lifecycle Governance

The operationalisation of dignity also requires reconsidering how military AI systems are reviewed over time. Existing legal obligations, particularly those associated with Article 36 of Additional Protocol I, provide essential mechanisms for evaluating new weapons, means, and methods of warfare – yet many AI systems differ from traditional military technologies in ways that complicate one-time review processes. As discussed above, software updates, retraining procedures, evolving datasets, and changing operational environments may alter system performance after deployment – a system assessed as reliable under one set of conditions may behave

---

[73] *The Sanremo Handbook on Rules of Engagement*, International Institute of Humanitarian Law, 2009, available at: https://iihl.org/wp-content/uploads/2022/12/ROE-HANDBOOK-ENGLISH.pdf.

differently under another. Dignitary design therefore supports the concept of lifecycle governance – rather than treating review as a singular event occurring prior to deployment, lifecycle governance recognises the need for continuous evaluation throughout the operational existence of a system. Monitoring, auditing, testing, incident review, and periodic reassessment should become essential components of humanitarian governance.[74] Such an approach does not replace Article 36 review – on the contrary, it builds upon its underlying logic: if States have an obligation to assess the legality of new military technologies before deployment, it is difficult to justify ignoring significant changes that may occur after deployment. Lifecycle governance therefore represents a natural extension of existing humanitarian obligations into the context of adaptive and data-driven systems.

VII. Conclusion

The emergence of artificial intelligence in military operations has generated intense debate concerning autonomy, accountability, and the future of warfare. Much of this discussion has focused on whether existing legal frameworks are capable of regulating increasingly sophisticated technologies. While these concerns remain important, this article has argued that they do not fully capture the challenge posed by military AI. The central question is not whether machines can become legal or moral actors, nor whether technological systems can simply be accommodated within existing regulatory structures. Rather, it is whether the human-centred normative commitments that underlie IHL can be preserved when the processes through which military decisions are generated become increasingly mediated by algorithmic systems.

The analysis has proceeded from the observation that contemporary military AI does not eliminate human agency but transforms the conditions under which agency is exercised. Decisions concerning the use of force increasingly emerge from interactions among commanders, operators, legal advisers, developers, procurement authorities, data curators, and algorithmic systems. Accordingly, operational cognition becomes distributed across complex socio-technical networks in which human actors remain formally responsible while relying upon informational structures that may be difficult to understand, challenge, or reconstruct. The challenge confronting IHL is therefore not the disappearance of responsibility but its reconfiguration under conditions of distributed agency.

At the same time, the article has sought to avoid overstating the novelty of technological change. Military decision-making has long relied upon technologies that mediate perception, communication, targeting, and operational planning. Artificial intelligence does not introduce technological mediation into warfare for the

[74] Alexandra T. Warner, *What is Monitoring in Humanitarian Action? Describing Practice and Identifying Challenges*, Active Learning Network for Accountability and Performance (ALNAP), London, 2017, pp. 11–14.

first time. Its significance lies in the increasing delegation of evaluative, predictive, and classificatory functions to systems that shape the informational environment within which humanitarian judgments are made. The legal challenge arises not because machines become moral actors but because algorithmic systems increasingly influence how human actors understand and evaluate the battlefield.

To address this challenge, the article has argued that human dignity should be understood as a structural principle of IHL. Although dignity occupies a less explicit position within humanitarian law than within international human rights law, it nevertheless informs the normative architecture of the humanitarian legal order. Through the principles of humanity, distinction, proportionality, precaution, and humane treatment, IHL affirms that persons affected by armed conflict retain an irreducible moral status that constrains the exercise of violence. Dignity functions, in part, as a safeguard against the instrumentalisation of human beings, ensuring that individuals are not reduced to mere objects through which military objectives are pursued. This understanding of dignity provides the foundation for the article's central contribution – the concept of dignitary design. Dignitary design is not proposed as a new regulatory mechanism, a substitute for meaningful human control, or an alternative to existing accountability frameworks. Rather, it is a jurisprudential framework for integrating respect for human dignity into the architecture, governance, evaluation, and operational deployment of military AI systems. It seeks to ensure that technological mediation does not erode the humanitarian commitments underlying IHL by embedding those commitments within the systems and institutions through which military decisions are increasingly shaped.

The article has further argued that preserving dignity in an era of military AI requires attention to three interconnected dimensions of governance. Architecturally, systems should be designed to support meaningful human judgment through explainability, transparency regarding uncertainty, safeguards against bias, and opportunities for human intervention. Procedurally, dignity requires ongoing review, auditing, testing, and lifecycle governance capable of addressing the evolving character of data-driven systems. Institutionally, it requires responsibility structures that reflect the distributed nature of contemporary military decision-making while ensuring that accountability remains anchored in human and institutional actors.

Operationally, dignitary design finds expression through proportionality assessments, precautionary measures, dignity-based auditing, rules of engagement, and continuous review mechanisms. These practices do not prohibit algorithmic assistance – rather, they seek to ensure that algorithmic systems remain subordinate to humanitarian judgment. The objective is not to prevent the use of AI in armed conflict but to preserve the capacity of military institutions to make decisions consistent with the values that humanitarian law exists to protect.

Ultimately, the challenge posed by military AI extends beyond questions of technological regulation – it concerns the future normative identity of IHL itself. Throughout its history, humanitarian law has adapted to changing means and methods of warfare while preserving its underlying commitment to humanity. Artificial intelligence presents a particularly significant test because it affects not only the instruments through which force is applied but also the processes through which judgments concerning force are formed. If humanitarian law is to remain a genuinely human-centred legal regime under such conditions, dignity cannot be treated solely as a retrospective standard for evaluating outcomes. It must also function as a principle shaping the systems through which decisions are generated.

The future of IHL will therefore depend not only on whether new technologies can be accommodated within existing legal frameworks but also on whether those frameworks continue to preserve human beings whose protection constitutes their fundamental purpose. In an increasingly algorithmic battlespace, the preservation of human dignity is not a peripheral ethical concern. It is a condition of maintaining legitimacy, coherence, and humanitarian character of the law itself. Dignitary design offers one possible path toward that objective by ensuring that, even as warfare becomes more technologically sophisticated, the normative centre of humanitarian law remains firmly human.

PART 3
GOVERNANCE

De Jure Necessity vs. De Facto Improbability? Establishing a New Treaty for Artificial Intelligence in the Military Domain

Zhixiong Huang[*]
Haokun Yu[**]

ABSTRACT

The military application of artificial intelligence challenges existing international law, which remains applicable but structurally insufficient to resolve emergent moral dilemmas, accountability gaps, and compliance deficits. Consequently, establishing a new treaty is a de jure necessity. However, this necessity confronts formidable de facto obstacles: the strategic imperatives of major military powers, prevailing scepticism towards treaty-making, and an unfavourable international political environment render the adoption of such an instrument almost de facto improbable. Nevertheless, this article argues that the treaty-making process can facilitate the formation of customary international law, serve as supplementary evidence for legal interpretation, and impose reputational costs on non-party states.

Keywords: AI in the military domain; Military AI; International humanitarian law; New treaty; International law-making process.

1. Introduction

Currently, the deployment of artificial intelligence (AI) in the military domain[1] is becoming a reality. In December 2025, the United Nations General Assembly once again noted the increasing integration of AI into military applications and highlighted its operational use.[2] Such international attention is well-founded. Reportedly, the United States deployed the Maven Smart System during the conflict

[*] Zhixiong Huang is the Hongyi Chair Professor at Wuhan University School of Law, and Director of the Wuhan University Institute for Cyber Governance. Email: fxyhzx@whu.edu.cn.
[**] Haokun Yu is a PhD Candidate in International Law at Wuhan University School of Law. Email: yuhaokun1012@whu.edu.cn.
[1] It should be noted that the terms "Military AI" or "AI in the military domain" often encompass a broad scope of applications, including weapon-related functions. Applications range from decision support systems for military operations to systems that support maritime security through counter-piracy efforts, counter-terrorism operations, and border security. In contrast, autonomous weapon systems (AWS) represent a partially specific application of such technology. While early multilateral discussions primarily focused on AWS, the international community has recently broadened its focus to address the broader implications of AI in warfare. Therefore, while the analysis in this paper remains largely relevant to AWS, the term "Military AI" is primarily used to reflect shifting international concerns and the potential expansion of future discussions. See UNIDIR, "Artificial intelligence in the military domain", New York, last visited 23 May 2026, available at: https://disarmament.unoda.org/en/our-work/emerging-challenges/artificial-intelligence-military-domain.
[2] UNGA Res. 80/58, 5 December 2025.

with Iran, while the Israel Defense Forces (IDF) utilized the Gospel and Lavender systems during the armed conflict in Gaza, all of which are military AI systems designed to enhance targeting capabilities.[3] These developments show that the regulation of military AI is no longer a potential consideration but has become an increasingly pertinent and contemporary issue. While AI's capabilities could offer significant military advantages in many areas, such as supporting military planning and human decision-making,[4] they also pose profound risks, including increasing the unpredictability of conflicts, accelerating escalation, and causing large-scale destruction.[5] These inherent risks necessitate robust regulation under international law.

In the realm of international law, academia and practitioners are deeply divided over whether a new treaty is necessary for military AI. Some countries and organizations believe that moral, accountability, and compliance issues compel states to adopt a new treaty to regulate it.[6] Meanwhile, other countries argue that the existing international law is sufficient to address these challenges.[7] They further contend that, without sufficient discussion regarding military AI, it is premature for states to establish a treaty to regulate autonomous weapon systems,[8] and they emphasize that military AI may have potential benefits, such as increasing the efficiency in task assignment and reducing the probability of errors.[9] This divide has become a focal point within the Convention on Certain Conventional Weapons (CCW) process and reflects the divergent positions of states.

---

[3] See Tara Copp *et al.*, "Anthropic's AI Tool Claude Central to U.S. Campaign in Iran, Amid a Bitter Feud", *The Washington Post*, 4 March 2026, available at: https://www.washingtonpost.com/technology/2026/03/04/anthropic-ai-iran-campaign/; Michael N. Schmitt, "Israel – Hamas 2024 Symposium – The Gospel, Lavender, and the Law of Armed Conflict", *Articles of War*, 28 June 2024, available at: https://lieber.westpoint.edu/gospel-lavender-law-armed-conflict/.

[4] Bill Boothby, "AI Warfare and the Law", *International Law Studies*, Vol. 104, 2025, p. 13.

[5] AutonomousWeapons.org, "The Risks of Autonomous Weapons", last visited on 23 January 2026, available at: https://autonomousweapons.org/the-risks/.

[6] According to Automated Decision Research, approximately 130 states currently support a legally binding instrument, including China, Germany, New Zealand, Switzerland, as well as member countries of the Non-Aligned Movement and the African Group. Conversely, only 12 states explicitly oppose such an instrument, including the United States, Russia, the United Kingdom, Japan, and India. Among international organizations, the ICRC has repeatedly called for the establishment of a legally binding instrument. See AutomatedDecisionResearch.org, "State Positions", last visited on 23 May 2026, https://automatedresearch.org/state-positions/?_state_position_negotiation=no (State Positions Database); ICRC, *ICRC Position on Autonomous Weapon Systems: ICRC Position and Background Paper*, Geneva, 12 May 2021. (ICRC Position)

[7] United States of America, *Submission of the United States of America*, 22 December 2023, p. 14, available at: https://docs-library.unoda.org/General_Assembly_First_Committee_-Seventy-Ninth_session_(2024)/78-241-US-EN.pdf.

[8] AutomatedDecisionResearch.org, "United Kingdom State Position", last visited on 23 May 2026, available at: https://automatedresearch.org/news/state_position/united-kingdom/.

[9] Russian Federation, *Working Paper of the Russian Federation "Application of the International Law to Lethal Autonomous Weapons Systems (LAWS)",* 18 July 2022, available at: https://documents.unoda.org/wp-content/uploads/2022/07/WP-Russian-Federation_EN.pdf.

While existing literature largely focuses on the analysis of legal grounds for establishing a new treaty for military AI (whether pro- or con-treaty),[10] it frequently overlooks the practical impediments to treaty formation. Such as limitations on arms control, inherent skepticism toward new legal instruments, and a fragmented international environment, all of which will significantly impact the treaty's success.[11] Meanwhile, the normative and factual value of the treaty-making process is also usually underestimated. Against this backdrop, this article aims to investigate the necessity and probability of establishing a new treaty for artificial intelligence in the military domain through a process-oriented perspective.[12] Based on the analysis of the challenges posed by military AI, Part 2 argues the necessity of establishing a treaty on a *de jure* basis. Part 3 then presents a realistic picture and discusses the *de facto* aspect, stressing the improbability of doing so due to several international political reasons. However, in Part 4, this article concludes on a cautiously optimistic view: the *de facto* improbability does not nullify the *de jure* necessity. In fact, the process of adopting a new treaty can serve as an incubator for customary international law (CIL), facilitate the interpretation of existing international law, and impose reputational costs on major military powers.

2. The De Jure Necessity: The Structural Insufficiency of Existing International Law

This Part examines the consensus and divergence between proponents and opponents of a new treaty from a legal perspective and emphasizes the necessity of the former. In fact, advocating for a new treaty rests not on the inapplicability of existing international law, but rather on its insufficiency. Specifically, the current international legal framework struggles to adequately address the moral dilemmas, accountability gaps, and compliance deficits created by military AI. Furthermore, the indeterminacy of applying existing rules to novel technologies undermines the legal certainty of international law, thereby diminishing the precondition for legal compliance. Collectively, these factors make the establishment of a new treaty *de jure* necessary.

---

[10] Benjamin Perrin, "Lethal Autonomous Weapon Systems & International Law: Growing Momentum Towards a New International Treaty", *ASIL Insights*, Vol. 29, No. 1, 2025; Laurie R. Blank, "New Treaty Law on Autonomous Weapons? An Opportunity to Reframe the Discourse", *Case Western Reserve Journal of International Law*, Vol. 57, No. 1, 2025, p. 233.

[11] For the purpose of this discussion, we define a successful treaty in its traditional sense as a legally binding instrument that secures the broad participation of major military powers and effectively binds their conduct in practice. However, the value of international regulation does not rely on achieving such an ideal outcome. Given its *de facto* improbability (as analyzed in Part 3), Part 4 argues that both the treaty-making process itself and a subsequent treaty with limited adherence would still generate independent value, regardless of the final instrument's formal success.

[12] Many legal theories, such as the New Haven school and International Legal Process schools, view international law as an ongoing, active process. In this study, we share this view. We believe that a treaty's impact goes beyond its written text. Instead, as countries continuously interact, the treaty can deeply influence how customary international law (CIL) is formed, how current laws are understood, and how political and reputational costs work outside the law. See Harold Hongju Koh, "Why Do Nations Obey International Law?", *The Yale Law Journal*, Vol. 106, No. 8, 1997, pp. 2618-2624.

2.1. The Applicability and Limits of Existing International Law

The assertion that a new treaty is required does not negate the applicability of existing international law. Although military AI can be considered novel technology, it does not preclude the applicability of the existing legal framework. This position finds support in the International Court of Justice's *Nuclear Weapon* advisory opinion. The Court held that although nuclear weapons were invented after most of the principles and rules of humanitarian law, "it cannot be concluded from this that the established principles and rules of humanitarian law applicable in armed conflict did not apply to nuclear weapons".[13] Otherwise, "such a conclusion would be incompatible with the intrinsically humanitarian character of the legal principles in question".[14] From this perspective, what international law primarily governs is the conduct of states and the consequences of their behaviour. Thus, while novel means and methods of warfare may generate unprecedented legal challenges, mere novelty of the tool does not preclude the general application of international law. This jurisprudential logic applies equally to military AI.

Furthermore, persuasive evidence of this applicability can be found in recent state practice. For example, within the United Nations framework, General Assembly Resolutions A/78/241, A/79/62, A/79/239, and A/80/58 all affirmed the applicability of international law, including the Charter of the United Nations, international humanitarian law (IHL), international human rights law, and international criminal law, to autonomous weapons systems and AI in the military domain.[15] Together, this jurisprudence and state practice unequivocally affirm the baseline applicability of existing international law.

What really divides proponents and opponents of a new treaty is not the applicability of existing international law, but its sufficiency. A typical argument against such a treaty is that existing international law could address the legal problems posed by new technologies. However, this article submits that such a view is overly optimistic. Although existing international law can be interpreted or even evolve, challenges posed by military AI are beyond its scope as military AI develops.

2.2. Moral Dilemmas

The first primary concern raised by military AI relates to the moral implications of its deployment. At the most fundamental level, opposition to Autonomous Weapons Systems (AWS) rests on the belief that it is morally unacceptable for machines to determine who lives or dies.[16] Compared with the AWS, humans have the moral

---

[13] International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, ICJ Reports, 8 July 1996, para. 86.
[14] *Ibid.*
[15] UNGA Res. 78/241, 28 December 2023; UNGA Res. 79/62, 10 December 2024; UNGA Res. 79/239, 31 December 2024; UNGA Res. 80/58, 5 December 2025.
[16] ICRC Position, p. 8.

agency to guide their actions.[17] Therefore, the moral elements may make people decide not to use force even if it would be lawful, whereas the AWS does not.[18] However, in the view of other scholars, robots can neither behave morally nor immorally.[19] Only human beings can be inhuman and choose not to comply with international law.[20] In their view, when military AI is well-regulated, it may help to facilitate compliance with international law and protect human life. More importantly, they argue that moral concerns do not automatically equate to legal obligations. While morally debatable, there are currently no positive rules explicitly prohibiting or restricting the use of military AI solely on moral grounds.[21]

Pragmatically speaking, this article admits that the application of military AI *per se* may not be immoral and should not be fully forbidden. After all, in the future, AI could perform better than humans in some situations regarding compliance with IHL.[22] In that case, applying military AI may promote compliance with international law rather than diminish its influence. However, under other circumstances, the moral risks posed by military AI can also not be overlooked. For instance, when an enemy combatant who is severely wounded and filled with terror fails to exhibit a standard surrender gesture due to fear or physical inability, a human soldier with empathy, morals, and instinct could sensitively perceive the genuine intention to surrender and withhold fire. In contrast, an algorithmically driven system lacking compassion might strictly adhere to its programming and engage the incapacitated target. Even though such an action may formally be consistent with the strict requirements of specific rules under IHL (such as the criteria for *hors de combat*),[23] it reduces the "live or dead" decision to a mere data-processing process. It is an action that puts human dignity aside and arbitrarily takes a human's life that originally could have been saved.

In this case, the use of military AI challenges the dictates of public conscience established by the Martens Clause of international law. The Martens Clause serves

---

[17] *Ibid.*

[18] *Ibid.*

[19] Marco Sassóli, "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified", *International Law Studies*, Vol. 90, 2014, p. 313.

[20] *Ibid.* at pp. 310, 313-314.

[21] See Jiawei Chu [褚家玮], *International Law and Autonomous Weapons* [国际法与自主武器], Current Affairs Press [时事出版社], Beijing, 2024, p. 167.

[22] See M. Sassóli, above note 19, p.320. Despite this, a counterargument is that the human/AI comparison reflects the assumption of static, hierarchical, and replaceable relations, as well as an oversimplification of the complex socio-technical interactions that characterise AI-enabled targeting practices. Therefore, AI technologies cannot be assessed *in abstracto*. On this point, see 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, Hague, 2026, p.52-70.

[23] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 41. (Additional Protocol I)

as a safety net for international law by requiring the application of principles of humanity and public conscience when specific regulations are absent.[24] However, the Martens Clause is inherently abstract, and applying it directly to regulate military AI remains difficult.[25] When specific positive rules are absent, and some specific means of applying military AI are deemed consistent with international law and could even bring military advantages, states are unlikely to refrain from using it solely on moral grounds. Because the Martens Clause primarily serves as a safety net rather than a detailed standard, the absence of a clear standard in existing international law leaves a dangerous legal vacuum. As aforementioned, the necessity and rationale for a new legally binding instrument is not that existing law is inherently flawed or inapplicable. It is because existing law is insufficient to echo the emerging new technological challenges. Thus, as the Martens Clause requires, a new treaty is necessary not just to debate morality but to translate these moral requirements into specific binding legal restrictions.

2.3. Accountability Gap

Another question raised by military AI is the accountability gap. Accountability demands that someone should be held liable for all wrong committed.[26] Thus, it is common to ask: when military AI has violated international law, like war crimes, crimes against humanity, or genocide, who should be responsible?[27] If the military AI commits all of the acts, should we allow AI to be the subject to take the responsibility? And is there any practical significance to doing this? In this regard, both Human Rights Watch and the Group of Governmental Experts (GGE) under the CCW framework have been concerned about the accountability issue raised by AWS.[28] For instance, the GGE has confirmed that an autonomous weapon system should not be anthropomorphized,[29] and IHL only imposes obligations on states, parties to armed conflict, and individuals, but does not impose obligations on machines.[30] Instruments as such seem to maintain the states' and individuals' status

---

[24] Under Additional Protocol I, the Martens Clause was expressed as follows: "In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity and from the dictates of public conscience." *Ibid.,* Art. 1 (2).

[25] For instance, Antonio Cassese contends that the Martens Clause is particularly ambiguous. It is very loosely worded and has consequently given rise to a multiplicity of often conflicting interpretations. Antonio Cassese, "The Martens Clause: Half a Loaf or Simply Pie in the Sky?", *European Journal of International Law*, Vol. 11, No. 1, 2000, pp. 187-188.

[26] Swati Malik, "Autonomous Weapon Systems: The Possibility and Probability of Accountability", *Wisconsin International Law Journal*, Vol. 35, No. 3, 2018, p. 620.

[27] See L. Blank, above note 10, at p. 234.

[28] Human Rights Watch, *Mind the Gap: The Lack of Accountability for Killer Robots*, New York, 9 April 2015; *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2018/3, 23 October 2018, para. 21(c).

[29] *Ibid.,* para. 21 (h).

[30] *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2019/3, 25 September 2019, para. 17(b). (GGE 2019 Report)

as the sole responsible entities. However, in practice, this issue remains controversial and complicates the practical allocation of responsibility for two primary reasons.[31]

First, the conduct and the consequences of the AWS are unpredictable.[32] Unlike conventional weapons, AWS is based on its trained algorithms rather than human direction, which means the human operator of the machine, in principle, cannot predict the future machine behaviour.[33] Because of this unpredictability, during military operations, even the human operator cannot fully predict the consequences of acts that violate international law. Thereby cutting the actual causal link between the human operator's actions and the AWS's outcomes, creating a significant accountability gap.[34] For instance, a troop deploys with AWS, whose accuracy is 99%, to conduct a mission aimed at destroying the enemy's armoured vehicles. However, during the mission, due to an algorithmic hallucination or an undetected problem, the AWS misrecognized a school bus as a target and conducted the attack.[35] Who should be responsible for that? If a human soldier committed such an action with intent or at his commander's direction, the soldier and his commander would certainly be held responsible under international criminal law. However, when such an action is conducted by an AWS, a lack of predictability fails to satisfy the strict mental element under international criminal law.[36] Specifically, the human commander does not satisfy the "should have known" standard under Article 28 of the Rome Statute,[37] nor can the operator of the AWS have "knowledge" under the "ordinary course of events" based on Article 30 in such a situation,[38] leaving both the soldiers and their commander unable to take the responsibility.

Second, during the development and operational use of AWS, "the problem of many hands" may also affect the distribution of responsibility.[39] This problem is relevant to the development and manufacturing of AWS, which requires the participation of programmers, engineers, different researchers, organizations, and companies.[40] When a mistake or malfunction occurs with AWS and violates

---

[31] It should be noted that the term "responsibility" addressed in the following paragraphs of this section does not refer to "state responsibility" reflected in the ARSIWA and CIL. Instead, it refers to individual responsibility, such as that of commanders, soldiers, programmers, and engineers.
[32] J. Chu, above note 21, at p. 171.
[33] See Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technology*, Vol. 6, 2004, p. 175.
[34] See J. Chu, above note 21, pp. 171-172.
[35] See Charles Trumbull, "Future of Warfare and Law Series – Addressing Uncertainty in the Use of Autonomous Weapons Systems", *Articles of War*, 5 November 2025, available at: https://lieber.westpoint.edu/addressing-uncertainty-use-autonomous-weapons-systems/.
[36] Jonathan Kwik, "Obstacles to Criminal Liability: A Systemic Analysis. In: Lawfully Using Autonomous Weapon Technologies", in *Lawfully Using Autonomous Weapon Technologies*, TMC Asser Press, Hague, 2024, pp. 328-334.
[37] Rome Statute of the International Criminal Court, 2187 UNTS 3, 17 July 1998 (entered into force 1 July 2002) (Rome Statute), Art. 28.
[38] Rome Statute, Art. 30.
[39] See J. Chu, above note 21, pp. 172-174.
[40] *Ibid*.

international law, it is usually difficult to isolate the responsibility of different participants.[41] Consider a scenario: a company with hundreds of software engineers and programmers developed an AWS with a set of flawed algorithms, but without any intent to commit a war crime, and such a system was extensively tested. Several years later, in an international armed conflict, the frontline commander may use and deploy such AWS in good faith and trust its verified accuracy in previous tests. However, in that real conflict, such a system has malfunctioned multiple times, causing unnecessary civilian casualties. Who should be responsible for that? Under international criminal law, such responsibility is inevitably diluted. Since the result was created by the engineer, programmer, and commander together, none of them, individually, could have violated international criminal law for lack of *mens rea*. Not to mention, after quite a long time, it may be difficult to trace which specific programmers or engineers made such a mistake and attribute the responsibility to them.

To address this accountability gap, many scholars have proposed various solutions, such as expanding the scope of responsibility under international criminal law,[42] establishing human judgment,[43] or establishing "war torts"[44]. However, all those views remain at the proposal stage; and none have been adopted as a rule of international law and have binding force on states and their behaviour. In fact, absent a dedicated treaty and broad international consensus, the realistic chance of these proposals being formally adopted into the current legal framework remains very slim. Admittedly, some elements among them, such as meaningful human control, could serve as an interpretative factor that guides the application of IHL. But at least in attributing criminal responsibility to individuals, such interpretative factors are grossly insufficient. Based on the principle of *nullum crimen sine lege,* rooted in international criminal law, the definition of a crime shall not be extended by analogy, and ambiguities shall be interpreted in favour of the person being prosecuted.[45] Therefore, merely interpreting existing laws to solve the accountability gaps would be fundamentally in conflict with the established legal framework. For instance, holding a commander or operator liable for unforeseeable violations caused by an AWS would require either extending the "knowledge" standard under Article 30 of the Rome Statute or expanding the scope of commanders' responsibility under Article 28. Both interpretations would be inconsistent with the rule of strict construction. Therefore, to hold humans accountable for AI's actions, the international community must establish new rules for the deployment of military AI.

---

[41] *Ibid.*

[42] Hin Yan Liu, "Refining Responsibility: Differentiating Two Types of Responsibility Issues Raised by Autonomous Weapons Systems", in Nehal Bhuta *et al.* (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, Cambridge, 2016, pp. 325-344.

[43] Jack M. Beard, "Autonomous Weapons and Human Responsibilities", *Georgetown Journal of International Law*, Vol. 45, No. 3, 2014, p. 621.

[44] See Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons", *University of Pennsylvania Law Review*, Vol. 164, No. 6, 2016, pp. 1347-1402.

[45] Rome Statute, Art. 22.

From this perspective, establishing a new treaty to discuss, enact, and confirm these principles and solutions is necessary.

2.4. Compliance Deficits

The absence of a dedicated treaty also creates significant practical challenges in complying with specific rules of IHL. The ambiguity of terms under existing international law and the uncertainty surrounding their application make compliance difficult and create numerous compliance deficits.

Concerning the legal review of new weapons, it is often posited that Article 36 of Additional Protocol I could serve as a mechanism to solve some of the humanitarian issues of military AI, as it requires the states to determine whether the weapon's employment would be prohibited by the Protocol or by any other applicable rules of international law.[46] It should be recognized that if the review process stipulated in Article 36 is strictly implemented, it is indeed possible to avoid the deployment of weapons that are incompatible with IHL. However, existing international law has not yet established a complete mechanism for the legal review of autonomous weapons systems.[47] It does not specify the scope of legal review, nor does it provide standardized substantive or procedural rules, leading to highly divergent state practices.[48] For instance, whether conventional weapons that are attached with artificial intelligence should be subjected to the weapon reviews is undetermined. Whether a uniform procedure should be applied during the review is uncertain. Whether the means and methods of warfare should be under review is undefined.[49] This uncertainty allows states to conduct reviews based on subjective preferences and varying standards, rendering Article 36 practically ineffective as a unified safeguard. Furthermore, in practice, most States Parties to Additional Protocol I may choose not to conduct formal weapon reviews.[50] If this persists, the role of Article 36 as a gatekeeper may be fundamentally undermined.

Admittedly, while Article 36 lacks specific procedural mechanisms for all weapons, this silence creates profound compliance problems when applied to military AI. For example, Article 36 triggers the review obligation upon the study, development, acquisition, or adoption of a "new weapon, means or method of warfare".[51] Since conventional weapons undergo virtually no substantive changes after deployment, states can meet this obligation by conducting a single weapons

---

[46] Additional Protocol I, Art. 36.
[47] Shouping Li [李寿平], "Improvement and Development of International Legal Regulation of Autonomous Weapon Systems" [自主武器系统国际法律规制的完善和发展], *Law Review* [法学评论], No. 1, 2021, p. 170.
[48] *Ibid.*
[49] *Ibid.*
[50] See James Farrant and Christopher M. Ford, "Autonomous Weapons and Weapon Reviews: The UK Second International Weapon Review Forum", *International Law Studies*, Vol. 93, 2017, p. 401.
[51] Additional Protocol I, Art. 36.

review prior to their operational use. However, for military AI, the adaptive capabilities gained through machine learning endow it with self-evolving characteristics.[52] This means it could potentially constitute a "new weapon" or "new method of warfare" at any given moment, necessitating weapon review throughout its entire lifecycle. Against this backdrop, conducting weapon reviews solely prior to deployment fails to meet the objective requirement for continuous scrutiny of military AI.[53] To ensure the weapon review mechanism really works and is effectively observed by states, a new treaty needs to refine the specific criteria for the review to provide effective guidance.

At the same time, the aforementioned Martens Clause, as one of the most fundamental principles of IHL, is also challenged by differing interpretations in its application. As early as 2012, Human Rights Watch and the International Human Rights Clinic at Harvard University argued that fully autonomous weapons would likely contravene the Martens Clause because they would run counter to the dictates of public conscience.[54] Despite such a broad interpretation being NGO-friendly, it is likely to be rejected by some States, as they prefer a narrow view of the Martens Clause and do not deem the principles of humanity or the dictates of public conscience as independent sources of international law.[55] Against the background of such differences, if a state merely regards the Martens Clause as a protective umbrella, not an independent source of obligation, then it cannot effectively guide the country's behaviour in military AI. Therefore, the advocacy of a new treaty does not necessarily require states to adopt a broader interpretation of the Martens Clause. Rather, it provides a forum for states to exercise their sovereign rights and translate abstract humanitarian concerns into concrete, negotiated treaty rules.

Both examples demonstrate that different interpretations of the same rules under international law can lead to varying levels of compliance. Some middle and small powers, as well as human rights organizations, are likely to adopt a strict standard to restrict major military powers and ensure that such military AI does not harm them. While some countries with strong military capabilities, because they are better able to deploy military AI in actual military conflict, are more likely to remain silent or take a loose interpretation of the rules meant to regulate them. Such a

---

[52] See Damian Copeland *et al.*, "The Utility of Weapons Reviews in Addressing Concerns Raised by Autonomous Weapon Systems", *Journal of Conflict and Security Law*, Vol. 28, No. 2, 2022, p. 294.

[53] As the rolling text notes, "any further development of, or modifications to, existing LAWS that have a significant impact on their target identification, selection, and engagement functions or on their anticipated effects, are reviewed to ensure compliance with IHL." However, such expressions currently do not have any legally binding force. See GGE on LAWS, "Rolling text", 18 December 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_2025.pdf.

[54] Human Rights Watch and International Human Rights Clinic, *Losing Humanity: The Case against Killer Robots*, United States of America, 19 November 2012, p. 4.

[55] See Tyler D. Evans, "At War with the Robots: Autonomous Weapon Systems and the Martens Clause", *Hofstra Law Review*, Vol. 41, No. 3, 2013, pp. 723-726.

deduction is evidenced by practice: while the Association of Southeast Asian Nations, [56] African Group, [57] and the Non-Aligned Movement [58] are actively proposing to establish a legally binding instrument, other nations, including the US, Russia, and Israel, emphasize the sufficiency of existing international law while simultaneously developing or deploying AWS in active conflicts. Thus, merely relying on existing international law will lead to multiple and fragmented interpretations of international law and further deepen divergences.[59] Without any specific action to facilitate the formation of an international consensus, the so-called "applying existing international law" approach will serve as a shield to prevent countries with strong military capabilities from being subject to international law.

2.5. Where Does the Necessity Come From: Legal Certainty as a Precondition for Compliance

From a jurisprudence perspective, legal certainty always plays a significant role in international legal systems.[60] Although the previous sections have identified specific doctrinal gaps, a more profound crisis lies in the loss of legal certainty. For hundreds of years, despite IHL possessing a degree of inherent constructive ambiguity, it has generally provided a stable and predictable baseline for states and offered a framework to guide them on what to do and what not to do during armed conflicts. However, in the context of rapidly developing military AI, the most basic clarity and predictability appear unattainable, and legal certainty has been eroded by legal indeterminacy.

---

[56] "Statement by His Excellency Chhea Keo Ambassador and Permanent Representative of the Kingdom of Cambodia to the United Nations, on behalf of the Association of Southeast Asian Nations (ASEAN)", ASEAN, for the 80th Session of the General Assembly, 23 October 2025, available at: https://reachingcriticalwill.org/images/documents/ Disarmament-fora/1com/1com25/statements/23Oct_ASEAN.pdf.

[57] "Statement delivered by Nigeria on behalf of the African Group Member States that are High Contracting Parties to Convention on Certain Conventional Weapons", November 2023, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional _Weapons_-Meeting_of_High_Contracting_Parties_(2023)/Nigeria_-_African_Group_-_2023_MHCP.pdf.

[58] Working paper by the Bolivarian Republic of Venezuela on behalf of the Non-Aligned Movement (NAM) and Other States Parties to the Convention on Certain Conventional Weapons (CCW), 2022, available at: https://documents.unoda.org/wp-content/ uploads/2022/08/WP-NAM.pdf.

[59] One might argue that if such divergences exist when states interpret existing international law, they may also exist when the international community is negotiating a new treaty. This is entirely possible due to states' differing considerations, which will be discussed in Part 3. However, there are still essential differences between establishing a new treaty and interpreting the existing law. Unilateral interpretation permits states to exploit the ambiguity of existing law to preserve their strategic flexibility, which, by its nature, favors states unwilling to regulate military AI. In contrast, the treaty-making process serves as a mechanism that compels states to articulate their legal positions and red lines. This mechanism will help build consensus within the international community and reduce divergences, as will be further discussed in Part 4.

[60] See James R. Maxeiner, "Some Realism About Legal Certainty in the Globalization of the Rule of Law", *Houston Journal of International Law*, Vol. 31, No. 1, 2008, p. 30.

Legal certainty in military AI is eroded by two forces, both of which make the mere interpretation of existing law incapable of solving the current legal predicament. Firstly, legal certainty is undermined by the fragmented interpretation. While states can give new meaning to existing rules to address technological changes, when multiple interpretations arise on the same issue, none of which has been widely recognized by the international community, there is no clarity in the law to follow.[61] Secondly, legal certainty is affected by the ambiguity of emerging soft law. When numerous proposals have been raised,[62] none of which has been confirmed as hard law, they cannot create any binding force for states. Meaningful human control is a significant example. It has been addressed in many diplomatic discourses but has not yet been incorporated into the *lex lata*.[63] In the absence of a clear treaty definition, states retain excessively broad discretion to interpret this concept, thus making "meaningful human control" lose its due limitation. Meanwhile, such instances also blur the line between legal obligation and political stance, increasing legal ambiguity and decreasing legal certainty. In this regard, merely relying on existing international law is insufficient. In an international society without a central legislative institution, existing law is neither able to actively choose which interpretation is authentic nor to adopt influential new proposals into it.

In this context, one might raise necessary concerns about the consequences of establishing a new treaty, and we must acknowledge their reasonableness. This perspective first contends that the pace of technological development far outstrips the speed of human imagination. Meaning, a new treaty may fail to address emerging issues in the future and create difficulties in defining regulatory content.[64] Second, this perspective also notes that a new treaty could undermine the existing protective framework of IHL.[65] This could lead to deviations in the protection afforded to specific actions from the provisions of the Geneva Conventions, or enable non-signatory states to argue that existing IHL does not apply to their military AI technologies.[66] In other words, such views hold that a new treaty not only fails to

---

[61] This phenomenon was partially reflected in the "Compliance Deficits" part in the former section.

[62] See, e.g., REAIM, "REAIM Pathways to Action", Summit on Responsible Artificial Intelligence in the Military Domain, Spain, 4-5 February 2026, available at: https://www.exteriores.gob.es/en/REAIM2026/Documents/REAIM%202026%20Pathways%20to%20Action.pdf; Lieber Institute for Law and Land Warfare, "Manual on International Law Applicable to Artificial Intelligence in Warfare", last visited 19 May 2026, available at: https://lieber.westpoint.edu/research/the-west-point-manual/.

[63] "However, the analysis of the concept of meaningful human control in this report shows that even if the concept is developing, it is not sufficiently defined to form the basis of a regulation on autonomous weapon systems at this point." See Amanda Musco Eklund, "Meaningful Human Control of Autonomous Weapon Systems: Definitions and Key Elements in the Light of International Humanitarian Law and International Human Rights Law", February 2020, p. 49, available at: https://www.fcas-forum.eu/ publications/ Meaningful-Human-Control-of-Autonomous-Weapon-Systems-Eklund.pdf.

[64] L. Blank, above note 10, pp. 251-252.

[65] *Ibid.,* p. 252.

[66] *Ibid.*

resolve future legal uncertainties but also introduces uncertainty in the application of existing law.

Regarding this perspective, this article acknowledges that these concerns pose challenges that must be addressed in the future treaty-making process. However, the article also contends that while these concerns seek to preserve legal stability, they inadvertently perpetuate legal uncertainty. First, the notion that "technological development outpaces human imagination" applies not only to new treaties but also to existing IHL. The aforementioned moral dilemmas, accountability gaps, and compliance deficits stem precisely from the states' failure to anticipate future technological developments in the past. Therefore, addressing pressing current issues based on the international community's present understanding of AI technology is far more meaningful than inaction. Simultaneously, this argument also misinterprets the purpose of establishing the treaties. Establishing new treaties does not equate to regulating specific, micro-level AI technologies. Rather, it involves creating international norms aligned with the overarching characteristics of AI technology, such as continuous weapons review and meaningful human control, which can also adapt to future technological transformations to a significant extent.

At the same time, concerns about undermining the existing framework for applying IHL overlook the relationship between general and special rules in international law. The establishment of a new treaty aims to apply IHL specifically in the context of military AI, not to replace general IHL. Even if there are concerns that this issue could serve as an excuse for some states to refuse to apply existing IHL, a new treaty could explicitly incorporate a "savings clause" stipulating that its application does not affect the application of existing IHL.

Based on this background, while existing international law still plays an important role in regulating military AI, establishing a new treaty in this field is *de jure* necessary. It is true that the establishment of a new treaty cannot eliminate all interpretation disputes, because international law is essentially the product of the coordination of will among States.[67] However, the purpose of establishing a new treaty is not to create a perfect regulatory mechanism overnight, but to clarify the necessary regulatory baseline, address emerging legal issues, and provide the most basic legal certainty. Firstly, a new treaty could offer an authoritative text that, while perhaps still containing compromises, could narrow the space for unilateral interpretation. Secondly, the treaty could actively consider emerging concepts and address new concerns, such as meaningful human control and moral dilemmas, thereby elevating soft law from *lex ferenda* to *lex lata*, filling legal gaps, and resolving legal problems. Only by doing so could states gain clear guidance and direction and comply with international law without ambiguity. Ultimately, the purpose of

[67] See Gengsheng Zhou (S. R. Chow) [周鲠生], *International Law* [国际法], The Commercial Press [商务印书馆], 2018, pp. 1-10.

establishing a new treaty is to use the precise text agreed upon by states to replace the legal interpretation that can exploit the ambiguity of the law, thereby achieving the rule of law in the field of military AI.

3. The De Facto Improbability: Political Impasse and Majority Follows Minority Dynamic

Although establishing a new treaty for military AI is *de jure* necessary, its establishment has been impeded by various factors. In this part, the article will examine the possibility of a treaty-making process under the CCW and any alternative process beyond it. This article submits that establishing a new treaty for military AI under the CCW mechanism is unlikely. Even if such a treaty were adopted in an alternative mechanism without the participation of the major military powers, it would be unlikely to effectively regulate the application of military AI. Having concluded that such a new treaty is *de facto* improbable, the article will outline the factors that drive this political and legal impasse and present a realist picture of the treaty-making process. Unlike most countries' expectations, the regulation of military AI is a process where the "majority follows the minority".

3.1. Possibility of Establishing a Military AI Treaty Under CCW and Alternative Mechanisms

3.1.1. CCW Mechanism

When considering the establishment of a new treaty for military AI, the CCW process is undoubtedly one of the most influential forums. Historically, this process has adopted five influential protocols regulating different types of conventional weapons ranging from non-detectable fragments and mines to incendiary weapons, laser weapons, and explosive remnants of war.[68] In 2013, the meeting of States Parties to the CCW decided that the meeting of 2014 would discuss the questions relating to lethal autonomous weapons systems (LAWS),[69] and would begin its regulatory process for AI in the military domain. Subsequently, the CCW has held annual meetings to discuss autonomous weapons systems,[70] and established a formal Open-Ended Group of Governmental Experts in 2016.[71] In 2019, the CCW

[68] 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, 1342 UNTS 137, 10 October 1980 (entered into force 2 December 1983).

[69] *Final Report of the 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*, UN Doc. CCW/MSP/2013/10, 16 December 2013, para. 32.

[70] The details of the discussed topic under the CCW mechanism can be found in B. Boothby, above note 4, pp. 405-472.

[71] *Final Document of the Fifth Review Conference 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,* UN Doc. CCW/CONF.V/10, 23 December 2016, p. 9.

mechanism published eleven guiding principles,[72] marking a significant milestone and a major consensus within the CCW at the time. However, such guiding principles are non-binding in nature, and since the start of the discussion regarding LAWS in 2013, there has been no mandate to negotiate a treaty.[73]

Contrary to initial optimism, many factors continue to block treaty negotiations under the CCW process. From the perspective of substantive differences in negotiations, countries have continued to diverge on definitions, technical, security, legal, ethical, and future aspects, none of which have been fully resolved.[74] Consequently, future negotiations are unlikely to proceed smoothly. Although the CCW process will remain the main forum for multilateral discussions on LAWS, international competition over the development and rule-setting power of LAWS will continue to intensify, and the prospects for LAWS arms control under the CCW process remain bleak.[75]

In fact, after the last CCW protocol was passed in 2003, more than two decades have passed without the adoption of a new protocol, and the future usefulness of the CCW process has been doubted in recent years[76]. The CCW process's weaknesses can be attributed to procedural requirements during new-protocol negotiations. The practice that has evolved during CCW negotiations is that consensus is required before adopting amendments to the Convention or Protocols, or adopting new Protocols.[77] However, with the growing number of CCW State Parties, achieving such a consensus is challenging.[78] During the current discussion regarding autonomous weapon systems, while about 130 states actively support the negotiation of a legally binding instrument, a small but decisive twelve states, including the U.S., Russia, India, Japan, South Korea, and Australia, explicitly refuse to adopt a legally binding instrument.[79] Despite their small number, under the

---

[72] See GGE 2019 Report, above note 30.

[73] See B. Boothby, above note 4, p. 406.

[74] The divergences under the CCW process include: (1) Definition Divergence: Should a feasible definition of LAWS be formulated now, and how should it be formulated? (2) Technical Divergence: What are the paths for exploring the characteristics of LAWS, and can strong artificial intelligence arrive, and when? (3) Security Divergence: Will the impact of LAWS be more beneficial or detrimental? (4) Legal Divergence: Does LAWS require new, additional regulation beyond international humanitarian law? (5) Ethical Divergence: Can machines become moral agents? Are machines tools for humans or intelligent agents? (6) Future Divergence: Should we implement preventative prohibition, temporarily allow it, or even encourage the development of LAWS? See Nengwu Xu [徐能武] and Kun Long [龙坤], "The Focus and Trend of Debate on Arms Control of Lethal Autonomous Weapons Systems under the UN CCW Framework" [联合国 CCW 框架下致命性自主武器系统军控辩争的焦点与趋势], *Journal of International Security Studies* [国际安全研究], No. 5, 2019, pp. 124-128.

[75] *Ibid.,* pp. 128-131.

[76] William H. Boothby, *Weapons and the Law of Armed Conflict*, 2nd ed., Oxford University Press, Oxford, 2016, p. 18.

[77] *Ibid.,* p. 99.

[78] *Ibid.,* p. 18.

[79] About the State Position Collection, see State Positions Database.

consensus requirements of the CCW, their refusal plays a veto role. Without a dramatic shift in position, the establishment of a new treaty through the CCW is improbable.

3.1.2. Alternative Mechanism Beyond the CCW

For the reasons above, the international community is also exploring alternative processes outside the CCW, aiming to negotiate a legally binding instrument. In 2022, the Human Rights Watch and the International Human Rights Clinic at Harvard Law School published a research report titled "An Agenda for Action: Alternative Processes for Negotiating a Killer Robots Treaty" to promote such an alternative mechanism.[80] The report suggested that the alternative mechanism should have the following characteristics: a common purpose of creating new law, a model of decision-making by majority vote, a clear and ambitious deadline, and an inclusive approach.[81]

When the CCW process is unworkable for negotiating a legally binding instrument, establishing a new treaty through an alternative process would be a smarter option and more likely to succeed. As Human Rights Watch observed, the Mine Ban Treaty, the Convention on Cluster Munitions, and the Treaty on the Prohibition of Nuclear Weapons all offer valuable precedents for establishing a treaty on military AI.[82] Taking the Convention on Cluster Munitions (CCM) as an instance: after consistently failing to reach a consensus under the CCW mechanism,[83] countries and organizations concluded that the existing negotiation framework was disproportionately influenced by major military powers, resulting in delays in developing a dedicated protocol on cluster munitions.[84] Therefore, after 2006, those countries and organizations that strongly advocated for a ban began seeking to establish an international treaty outside the framework of the CCW.[85] After the launch of the Oslo Process in February 2007, they finally reached a Convention in 2008.[86] Compared to the long, fruitless CCW negotiation process, the Oslo Process took only about 2 years and resulted in a treaty. The advantages of the alternative

[80] See Human Rights Watch and International Human Rights Clinic, *Agenda for Action: Alternative Processes for Negotiating a Killer Robots Treaty*, United States of America, November 2022.
[81] *Ibid.*, p. 4.
[82] *Ibid.*, p. 15-25.
[83] Zhixiong Huang [黄志雄] and Lin Fan [范琳], "The 2008 Convention on Cluster Munitions under the Trend of Humanization of International Law" [国际法人本化趋势下的 2008 年《集束弹药公约》], *Law Review* [法学评论], No. 1, 2010, p. 69.
[84] *Ibid.*
[85] *Ibid.*
[86] Zhixiong Huang [黄志雄], "Non-Governmental Organizations and the Convention on Cluster Munitions in 2008: Participation, Influence and Revelation" [非政府组织与 2008 年《集束弹药公约》：参与、影响及启示], *Wuhan University International Law Review* [武大国际法评论], Vol. 13, 2008, pp. 351-352.

mechanism stand out. Compared to previous successful experiences, the regulation of military AI faces a similar situation. Both scenarios enjoy sufficient support for such a treaty. Regarding AWS regulation, approximately 130 countries support the establishment of a legally binding instrument.[87] Therefore, despite the blocking elements, achieving a binding treaty under an alternative mechanism in the long term is technically feasible.

However, feasibility must not be confused with the treaty's success. Notwithstanding the possibility of a new treaty being adopted, using such a treaty *per se* to effectively regulate the international community is another question. The fundamental problem of an alternative mechanism is that the key stakeholders of military AI may be absent. Historical precedent on nuclear weapon control provides stark evidence on this issue. For instance, despite ninety-five States having signed the Treaty on the Prohibition of Nuclear Weapons,[88] none of the recognized nuclear-weapon states have participated in it. Instead, they firmly and directly state their opposition to the Treaty on the Prohibition of Nuclear Weapons and reaffirm the application of the Treaty on the Non-Proliferation of Nuclear Weapons.[89] This precedent reminds us that, when the same logic applies to the regulation of military AI, the countries that are most likely to own and use AWS are also highly unlikely to participate in a potential convention on military AI. If this is the case, only countries that do not possess AWS would be bound by the treaty, while those that control the technology would operate without restriction; this outcome would render the convention *de facto* futile.

3.2. The Underlying Logic of Weapons Regulation and Treaty-Making

Several factors explain the *de facto* improbability of establishing the treaty. This section examines these obstacles from three perspectives: the strategic imperatives of major military powers, prevailing scepticism towards treaty-making among Western countries, and the current unfavourable international order and political environment.

3.2.1. The Strategic Imperatives of Major Military Powers

From a historical perspective, Hans J. Morgenthau has elaborated on the limitations and the fate of arms control. As an effort to strengthen international peace by increasing military stability, it has failed repeatedly with conventional weapons and

[87] State Positions Database, above note 6.
[88] UNODA, "Treaty on the Prohibition of Nuclear Weapons, Participants", last visited 24 May 2026, available at: https://treaties.unoda.org/t/tpnw/participants.
[89] The Ministry of Foreign Affairs of the Russian Federation, "Joint Statement by China, France, Russian Federation, United Kingdom and United States", 29 October 2018, available at: https://www.mid.ru/en/foreign_policy/news/1577389/.

has succeeded only in the nuclear field.[90] Such control of nuclear weapons is made possible in theory because the ability of major nuclear powers has reached its optimum of assured destruction, beyond which it is irrational to go.[91] By contrast, in conventional realms, it has failed because the quantity and disposition of such weapons will directly affect the distribution of military power and military advantage, thereby undermining effective regulation.[92]

Artificial intelligence, as a civil-military dual-use technology, exhibits characteristics closer to those of conventional weapons than to those of nuclear weapons. Compared to nuclear weapons that guaranteed deterrence and mutual assured destruction, military AI can provide sophisticated decision support and conduct precision strikes directly. These characteristics mean that the deployment of military AI directly dictates battlefield supremacy. Given its disruptive potential to alter the balance of power, states are locked in a security dilemma.

Unwillingness to establish a treaty in related fields is by no means unique to one country. On the contrary, it reflects the strategic considerations among all technologically advanced countries. For example, although China has declared its support for the establishment of a new treaty,[93] the structure of the international system determines that, without the commitment of its military adversaries, no great power, including China, Russia, and the US, will give up such a significant military advantage and join such a treaty directly. Instead, all of these countries are integrating AI into their militaries and trying to significantly increase their military advantages.[94] Under this situation, if a group of states, such as the US and its allies,

---

[90] Hans J. Morgenthau, *Politics among Nations: The Struggle for Power and Peace*, 7th ed., revised by Kenneth W. Thompson and W. David Clinton, McGraw-Hill/Irwin, New York, 2005, p. 423.
[91] *Ibid.*
[92] *Ibid.*
[93] Indeed, such support for establishing a binding instrument was not unconditional. As Ambassador Shen Jian stated regarding the facilitation of a legally binding instrument: "It deserves further discussions on these specific matters. It is recommended that the GGE further elaborate the interpretation, application, and implementation of specific rules of IHL, taking into consideration the nature of emerging military technologies, their modes of operation, and the operational environment." See UNODA, "Ambassador SHEN Jian's Remarks at the First Session of the 2024 CCW Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapon Systems (GGE on LAWS)", 2024, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/GGE_LAWS_-_China_-_General_statement_-_English_Copy.pdf; Permanent Mission of the People's Republic of China to the UN Office at Geneva and Other International Organizations in Switzerland, "Statement by Ambassador SHEN Jian at the Second Session of the 2024 CCW Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapon Systems", 27 August 2024, available at: https://geneva.china-mission.gov.cn/eng/dbtxwx/202408/t20240827_11480159.htm.
[94] Priyesh Mishra, *et al.*, *Code, Command, and Conflict: Charting the Future of Military AI*, Harvard Kennedy School Belfer Center for Science and International Affairs, Cambridge, MA, 12 December 2025, pp. 5-6, available at: https://www.belfercenter.org/research-analysis/code-command-and-conflict-charting-future-military-ai.

explicitly refuses to participate in this treaty, it would be extremely dangerous for their adversaries, such as China and Russia, to accept such unilateral constraints.

Consequently, applying Morgenthau's explanation to military AI could help explain why several countries are unwilling to participate in its regulation. On the one hand, since none of the states can ascertain their adversaries' strategic intentions, developing military AI that could enhance their own capabilities, rather than restricting them, becomes the only viable solution for some states. In this context, unilateral constraints will only increase the vulnerability of states. Meanwhile, for nations leading in military capabilities, the motivation behind developing military AI extends beyond mere strategic defence. Instead, they seek to cultivate an absolute advantage through asymmetric means and widen the gap in military superiority over other nations. Therefore, the "security dilemma" and the pursuit of absolute hegemony still exist in the era of AI. Without substantial changes to the global security situation, the development, deployment, and application of military AI as a strategic imperative for major military powers are unlikely to be directly constrained by treaty law.

3.2.2. Scepticism Towards Treaty-Making among Western Countries

Another reason for the treaty remaining improbable may stem from several Western countries' scepticism toward treaty-making in the technology realm. Such a sceptical stance is primarily held by some of the technologically advanced states led by the US and the UK.[95] Beyond the realm of military AI, such scepticism was also evident in cyberspace governance.[96] On the surface, such scepticism stems from the concern that slowly evolving international rules are unable to keep pace with rapid

[95] The countries mentioned here primarily include Australia, Estonia, Israel, Japan, Poland, the Republic of Korea, the United Kingdom, and the United States.

[96] For example, in 2011, the US vice president expressed his opposition to the establishment of an international legal instrument and stressed the applicability of existing rules; the same view was also reflected in the U.S. "International Strategy of Cyberspace" of that year. See The White House, "VP's Remarks to London Cyberspace Conference", 1 November 2011, available at: https://obamawhitehouse.archives.gov/the-press-office/2011/11/01/vps-remarks-london-cyberspace-conference; The White House, *International Strategy for Cyberspace: Prosperity, Security, and Openness in a Networked World*, Washington, DC, 2011, p. 9.

technological change.[97] To avoid such consequences, some Western countries argue for the interpretation of existing international law.[98]

However, at a deeper level, such scepticism is rooted in the contestation over international rule-making power among countries. To understand this logic, one can draw a direct parallel to the historical struggles in the negotiation process of international treaties on cybercrime. Historically, major Western powers promoted the application of the Budapest Convention, spearheaded by the Council of Europe,[99] and took a negative attitude towards the UN Convention Against Cybercrime, which involved deep participation from China, Russia, and the global south. The basic attitude of Western countries toward the Convention against Cybercrime can be reflected in their diplomatic records. For instance, the US deputy ambassador commented that the resolution on the Convention against Cybercrime will "undermine international cooperation to combat cybercrime".[100] The EU also announced that the Convention against Cybercrime will "lower the standards for protecting human rights and fundamental freedoms, increase the digital divide, and endorse state control of the Internet".[101] On the contrary, despite the global south countries also having internal differences on specific terms of the treaty, these supporters advocate for a more inclusive global process. This divergence reflects that countries have a stronger preference for the framework in which they are deeply involved. The preferences for the existing framework on military AI also reflect Western countries' desire to shape international rules through existing platforms and interpretive methods, where they have historically held dominant influence.

At the same time, on a practical level, this scepticism toward new treaties enables Western states to steer the evolution of existing international norms in a

---

[97] In 2013, the United Kingdom expressed its view that "The United Kingdom does not believe that attempts to conclude comprehensive multilateral treaties, codes of conduct or similar instruments would make a positive contribution to enhanced cybersecurity for the foreseeable future. The complex and comprehensive nature of any binding agreement across the entirety of a cyberspace that is evolving at "net speed" means that it could not be effective or command widespread support without many years, possibly decades, of painstaking work on norms of behaviour and confidence-building measures to build up the necessary understanding and trust among signatories and to ensure that they can be reliably held to account for their adherence to their commitments." See UN General Assembly, *Developments in the field of information and telecommunications in the context of international security, Report of the Secretary-General*, A/68/156, 16 July 2013, pp. 18-19.

[98] The UK further noted that "the efforts of the international community should be focused on developing common understandings on international law and norms rather than negotiating binding instruments that would only lead to the partial and premature imposition of an approach to a domain that is currently too immature to support it." See *Ibid.*

[99] See Delegation of the European Union to the Council of Europe, "EU Statement in support of the Council of Europe Convention on Cybercrime", 15 January 2020, available at: https://www.eeas.europa.eu/delegations/council-europe/eu-statement-support-coun cil-europe-convention-cybercrime_en?s=51. (EU Statement)

[100] See Edith M. Lederer, "UN Gives Green Light to Draft Treaty to Combat Cybercrime", *The Associated Press*, 28 December 2019, available at: https://apnews.com/general-news-79c7986478e5f455f2b281b5c9ed2d15.

[101] See EU Statement, above note 99.

direction favourable to their interests by championing soft law norms. For instance, in the realm of cyberspace governance, the currently established "norms of responsible state behaviour in Cyberspace" largely reflect the strategic priorities of Western states.[102] In contrast, issues of critical concern to many non-Western nations, such as cyber terrorism and cyber surveillance, are virtually absent from these existing norms.[103]

Therefore, beyond the explicit justifications offered in national position statements, Western countries' scepticism towards new treaties can be seen as a strategy to leverage their dominance in the international rule-making process. The ultimate goal is to shape the future development of existing international law. Regarding the question of whether a new treaty should be established for military AI, it is crucial to recognize the underlying reality: the scepticism maintained by Western and certain other states is, in essence, a manoeuvre to utilize alternative mechanisms to guide the evolution of international law toward their preferred outcomes.

However, it should also be admitted that, unlike in cyberspace governance, divergences exist in the approach to regulating military AI within Western countries. Despite countries like Germany, France, Canada, and Austria adopting a strategy similar to that of the US and UK in cyberspace governance, they are more active in supporting a legally binding instrument or taking a more prudent view on military AI[104] instead of merely opposing a new treaty. This situation likely arises because cyber operations normally occur below the threshold of use of force, and European countries are hence more likely to value their strategic flexibility. In contrast, military AI is directly relevant to lethal attacks and human lives, triggering the red line of IHL directly. Therefore, for countries like France and Germany, the concern for basic humanitarian principles outweighs the concern for flexible regulation. As these countries stressed in 2022, the international community should take a "two-tier approach" within the UN framework, prohibit autonomous lethal weapons outside human control, while strictly regulating others to ensure meaningful human control.[105] Meanwhile, Germany even directly supported the establishment of a new treaty.[106] Therefore, although the overall prospect of establishing a new treaty remains uncertain, such internal divergences objectively increase the chances of its successful establishment.

---

[102] Zhixiong Huang [黄志雄], "Norms of Responsible State Behavior in Cyberspace: Origins, Impacts and Responses" [网络空间负责任国家行为规范：源起、影响和应对], *Contemporary Law Review* [当代法学], No. 1, 2019, pp. 67-68.

[103] Of course, it cannot be denied that the norms of responsible state behavior in cyberspace themselves possess considerable rationality and merit. Moreover, the broad scope and non-binding nature of such norms make non-Western countries like China and Russia willing to participate in them. See *Ibid.*

[104] See State Positions Database, above note 6.

[105] See *Working paper Submitted by Finland, France, Germany, the Netherlands, Norway, Spain and Sweden*, UN Doc. CCW/GGE.1/2022/WP.7, 9 August 2022.

[106] State Positions Database, Germany, above note 6.

3.2.3. Current Unfavourable International Order and Political Environment

Third, the prevailing international order and political environment are unfavourable for the establishment of a new treaty. From an international political perspective, the intensification of geopolitical rivalry has substantially shaped an environment that incentivizes the deployment of AI weaponry rather than its restriction. Geopolitically, the international arena is currently defined by active regional "hot wars" involving the use of military AI. In the Russia-Ukraine conflict, both parties have leveraged AI-enabled drones and decision-support systems to gain significant battlefield advantages.[107] Similarly, in the Israel-Palestine conflict, the "Gospel" system has significantly accelerated the generation of targets for the Israeli military.[108] Furthermore, during recent US military operations against Iran, military AI systems such as Project Maven have been extensively deployed to identify targets.[109] The demonstrated operational utility of such applications is likely to cause states facing potential future conflicts to adopt a prudent approach toward any restrictions on military AI.

From a macro perspective, the contemporary international environment also no longer provides a foundation for establishing arms control treaties. The erosion of the rule-based international order by the great powers in the past decade has undermined the foundation of international cooperation. For example, as a dominant global actor, the United States' recent disregard for international norms has precipitated a degradation of the legal order, ushering the international community into an era akin to the "collapse of rites and music" (*Li Beng Yue Huai*, 礼崩乐坏).[110] Since the inauguration of the Trump Administration in 2025, a series of flagrant violations of international law, ranging from the detention of the

[107] Abdujalil Abdurasulov, "The New AI Arms Race Changing the War in Ukraine", *BBC*, 10 October 2025, available at: https://www.bbc.com/news/articles/cly7jrez2jno.

[108] Harry Davies, et al., "'The Gospel': how Israel uses AI to select bombing targets in Gaza", *The Guardian*, 1 December 2023, available at: https://www.theguardian.com/ world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

[109] Michael Brown, "The First AI War: How the Iran Conflict Is Reshaping Warfare", *Forbes*, 30 March 2026, available at: https://www.forbes.com/sites/mikebrown/2026/03/30/the-first-ai-war-how-the-iran-conflict-is-reshaping-warfare/.

[110] Derived from The Analects, the concept of "the collapse of rites and music" (*Li Beng Yue Huai*, 礼崩乐坏) describes a state of social disorder and normative decay. During China's Zhou dynasty (1046–256 BCE), waning quasi-central authority caused established ritual institutions to lose their binding force. Today, although lacking a central authority comparable to the Zhou court, the international community relies on international law as its normative framework. As the efficacy of international law declines, the contemporary global order increasingly mirrors this historical period of "the collapse of rites and music".

Venezuelan President [111] to assertions of sovereignty over Greenland, [112] have triggered widespread scepticism regarding the "rules-based international order" and the very efficacy of international law. As Canadian Prime Minister Carney warned at the 2026 World Economic Forum: "the rules-based order is fading... the strong do what they can, and the weak must suffer what they must."[113] As the international community increasingly shifts from idealism to realism, the efficacy of international law—inherently devoid of coercive enforcement—becomes progressively limited.

It should be admitted that some new multilateral treaties have been negotiated in recent years. The Convention against Cybercrime is a prominent one. Facing widespread opposition, the treaty still passed in 2024 and was opened for signature in 2025. At first glance, such an instance seems to provide the possibility to establish a new treaty in an unfriendly environment. But in essence, the Convention against Cybercrime is a treaty that authorizes states to cooperate against non-state actors, while the potential treaty on military AI might constrain state power. When states' perception of international relations becomes more realist and cooperation on enforcement might exist in the future, requiring states to unilaterally disarm would be more difficult. Consequently, attempts to employ international law to constrain state power have become increasingly difficult, rendering a new treaty on military AI *de facto* improbable.

3.3. The "Majority Follows Minority" Dynamic in Military AI

Under the cumulative impact of the aforementioned factors, despite the sustained efforts of the international community, the prospect of establishing a treaty or achieving effective regulation through one remains bleak. This reality presents a unique landscape in the regulation of military AI: while the vast majority of states support a binding treaty, the opposition of a small minority of countries renders such a treaty improbable to conclude or incapable of having substantive legal force. In this sense, the regulation of military AI has devolved into a process where "the majority follows the minority".

Within the CCW framework, the requirement for consensus nominally grants every state the right to have its will respected. However, in substance, it accords a *de facto* veto to major military powers capable of influencing the outcomes of progress. Under alternative mechanisms, even if a treaty secures widespread

---

[111] Michael N. Schmitt, et al., "International Law and the US Military and Law Enforcement Operations in Venezuela", *Just Security*, 5 January 2026, available at: https://www.justsecurity.org/127981/international-law-venezuela-maduro/.

[112] OHCHR, "Greenland: UN Experts Urge United States to Respect International Law and Right to Self-Determination", 14 January 2026, available at: https://www. ohchr.org/en/press-releases/2026/01/greenland-un-experts-urge-united-states-respect-international-law-and-right.

[113] Mark Carney, "Davos 2026: Special address by Mark Carney, Prime Minister of Canada", *World Economic Forum*, 20 January 2026, available at: https://www.weforum.org/stories/2026/01/davos-2026-special-address-by-mark-carney-prime-minister-of-canada/.

participation, it represents merely a "symbolic victory". As long as the few states possessing actual military AI capabilities refuse to join, the treaty's capacity to alter battlefield realities will approach zero. Consequently, the influence of major military powers over the regulation of military AI remains disproportionately immense. In the absence of binding hard law, middle and small states eager to regulate military AI are constrained by these major powers and may be compelled to accept soft-law norms imposed by them. In this regard, the United States-led Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy is a prominent example.[114] On the one hand, the US explicitly refuses to establish a legally binding new treaty. On the other hand, it is trying to adopt a political declaration to transfer its national will into normative proposals to increase its rule-making power in the military AI, and call on more than fifty middle and small states to endorse it.[115] This allows major powers to seize the agenda-setting power and ensure the field remains under their dominance.

In this context, the "majority follows minority" dynamic underscores that state power remains the determining factor in whether a treaty can be established and effectively implemented. If the broader community of nations seeks to reclaim the initiative in the regulatory game and surmount the obstacles posed by the minority, they must adopt strategic approaches. This will be the focus of the discussion in Part 4.

4. From De Jure Pursuit to De Facto Law: The Treaty Process as an Incubator, Evidence, and Reputational Costs

Despite the improbability of establishing a treaty to effectively regulate military AI, this does not render the pursuit of such an instrument futile. In fact, this article adopts a cautiously optimistic view regarding the normative value of the treaty-making process. In a landscape defined by the "majority follows minority" dynamic, the majority of states can leverage the treaty-making process to reclaim the initiative in the international rule-making process through three methods. First, the treaty-making process itself could serve as an incubator for the formation of CIL. Second, the treaty text could serve as substantive evidence influencing the interpretation of existing international law. Finally, the international community could impose constraints through legal and reputational costs on non-party states. Through these approaches, both treaty participants and non-participants will be effectively regulated within the evolving regulatory framework for military AI.

---

[114] US Department of State, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy", 2023, https://www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy. (Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy)

[115] *Ibid.*

4.1. The Treaty as an Incubator: Facilitating the formation of CIL

The treaty-making process may facilitate the formation of CIL. Admittedly, grounded in the fundamental rule that *pacta tertiis nec nocent nec prosunt*, a treaty cannot impose direct binding obligations on non-party states.[116] However, this does not imply that such treaties are devoid of influence on non-parties. According to Article 38 of the Vienna Convention on the Law of Treaties (VCLT), treaty provisions reflecting CIL remain binding on all states in that capacity.[117] Under specific circumstances, treaty provisions may even transform into general international law.[118] Through this mechanism, the core norms under the military AI treaty, like meaningful human control, have the potential to generate or crystallize into CIL and exert universal binding force upon non-party states.

The formation of CIL depends on two fundamental elements: state practice and *opinio juris*. In terms of state practice, as the ICJ elaborated in the *Asylum* Case, constant and uniform usage is necessary for the formation of CIL.[119] Meanwhile, *opinio juris* requires that such practice was an act in a way that they believed themselves legally compelled to do so.[120] These requirements dictate that the emergence of CIL in the field of military AI must go beyond the mere political behaviour, and must demonstrate widespread, representative, and consistent practice.[121] In this regard, the treaty-making process and the resulting text could promote the development of CIL in two ways.

First, the treaty-making process would facilitate the states to articulate their legal positions on military AI regulation, thus generating relevant state practice with *opinio juris*. To date, the international consensus in this field is still limited to political

---

[116] Vienna Convention on the Law of Treaties, 1155 UNTS 331, 23 May 1969 (entered into force 27 January 1980), Art. 34. (Vienna Convention on the Law of Treaties)

[117] Vienna Convention on the Law of Treaties, above note 116, Art. 38.

[118] As the ICJ elaborated in the *North Sea Continental Shelf* Case: "A norm-creating provision which has constituted the foundation of, or has generated a rule which, while only conventional or contractual in its origin, has since passed into the general corpus of international law, and is now accepted as such by the *opinio juris*, so as to have become binding even for countries which have never, and do not, become parties to the Convention. There is no doubt that this process is a perfectly possible one and does from time to time occur: it constitutes indeed one of the recognized methods by which new rules of customary international law may be formed." See ICJ, *North Sea Continental Shelf Cases (Federal Republic of Germany v. Denmark; see Federal Republic of Germany v. Netherlands)*, Judgment, ICJ Report 1969, para. 71 (ICJ, *North Sea Continental Shelf Cases*); see Giorgio Gaja, "1969 Vienna Convention Article 38 Rules in a treaty becoming binding on third States through international custom", in Olivier Corten *et al*. (eds), *The Vienna Conventions on the Law of Treaties: A Commentary*, 2nd ed., Oxford University Press, Oxford, 2011, p. 954-956.

[119] International Court of Justice, *Colombian-Peruvian Asylum Case*, Judgment, ICJ Report 1950, pp. 276-277.

[120] Malcolm N. Shaw, *International Law*, 9th ed., Cambridge University Press, Cambridge, 2021, p. 74.

[121] *Report of the International Law Commission of its Seventieth Session*, U.N. Doc. A/73/10, p. 120, Conclusion 8 (2018). (ILC 2018 Report)

declarations[122] and lacks the "accepted as law" characteristic, thereby obstructing the formation of CIL. Although CIL may gradually develop in the absence of treaties, this process will undoubtedly be slow, given the prevailing caution among states.[123] In this context, the treaty-making process would accelerate CIL's generation and crystallization. In fact, the treaty-making process or draft treaty provisions may themselves give rise to State practice and evidence of acceptance as law.[124] If the international community can effectively leverage this mechanism and initiate the treaty negotiation process, the textual formulations negotiated by states during treaty negotiation can become state practice reflecting their *opinio juris*, thereby transcending mere political commitments. By providing multilateral forums for states to articulate their practices and *opinio juris*, the treaty-making process offers a practical pathway for the eventual formation of CIL.

Secondly, the treaty-making process could pressure minority opposing states to publicly articulate their legal positions, reducing the strategic ambiguity that they intentionally maintain and facilitating the generation and crystallization[125] of CIL. Under the current framework, AI is regulated solely through the interpretation of existing laws. If a nation wishes to avoid international oversight, it need only remain silent at no cost or employ strategic ambiguity regarding specific legal standards to achieve its objectives. For instance, for rules such as "meaningful human control", a state may repeatedly affirm its applicability while avoiding clarification of the specific content. However, the formal treaty-making process would fundamentally change this phenomenon. During the negotiation of formal treaty provisions, each article drafted may be relevant to future state obligations and state responsibility. If states completely refuse to participate in the process of expressing legal positions both within and outside the treaty process, once specific treaty provisions crystallize into CIL, their silence may be regarded as implied consent, thereby compelling them to accept the rules they opposed. Therefore, if a draft treaty explicitly requires AWS to be subject to human control before launching each specific attack, opposing states would have to express their legal objections to protect their military advantages and clearly state the legal standards they support. Thereby helping to establish a "baseline consensus" on military AI within the international community and serving as the starting point for the generation of CIL.

---

[122] Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, above note 114.

[123] In fact, the formation of CIL is inherently time-consuming. Under the best of circumstances, CIL may be created in as little as 10 to 15 years. See Matthijs Maas, "International Law Does Not Compute: Artificial Intelligence and the Development, Displacement or Destruction of the Global Legal Order", *Melbourne Journal of International Law*, Vol. 20, No. 1, 2019, p. 50.

[124] ILC 2018 Report, above note 121, p. 143, Commentary to Conclusion 11, para. 1.

[125] Crystallization means that the custom was *in statu nascendi* when the treaty was drafted, but became a customary rule subsequently. See Marina Fortuna, "Interpretation of Customary Rules by Reference to Treaties and General Principles of Law", in Marina Fortuna *et al*. (eds), *Customary International Law and Its Interpretation by International Courts: Theories, Methods and Interactions*, Cambridge University Press, Cambridge, 2024, pp. 269-270.

Nevertheless, the transformation of treaty provisions into CIL is subject to several important limitations. During this process, while the treaty accepted by a sufficient number of states is required,[126] it should also include all states that have the opportunity and possibility to apply the rule in question.[127] In particular, the participation of "specially affected states" remains crucial in this process.[128] If this holds true, the subsequent practice of non-party states could substantially affect the formation of CIL. If the specially affected opposing states, such as the US, Russia, and the UK, continue to oppose the provisions of a new treaty, the formation of CIL may face significant obstacles. However, building on the treaty's broad international support, and constrained by considerations of international reputation and concerns over the treaty's stigmatizing effects, they may opt to align their practices with the treaty's provisions in practice.[129] In such circumstances, the treaty's provisions themselves will be widely emulated by states, thereby facilitating the formation of CIL.[130] Even if these specially affected states persistently object to the formation of the CIL, the accumulated state practices and *opinio juris* under the treaty would still lay a robust foundation for future normative development. More importantly, even if these countries cross the red line drawn by the treaty in practice, the treaty ensures they can no longer do so under the guise of legal ambiguity, thereby forcing them to bear the full political responsibility of their actions. Therefore, rather than directly generating CIL, the treaty-making process may function as an incubator that provides a suitable environment and accelerates the formation of state practice and *opinio juris*.

Ultimately, while acknowledging the complexity of the issue, a forward-looking perspective must be maintained in regulating military AI. Historically, the establishment of international legal regimes governing domains such as warfare, navigation, trade, and human rights has spanned decades or even centuries, evolving into the frameworks we have today through long-term development. In contrast, the regulation of military AI, since the CCW mechanism's intervention in 2013, has a history of only a bit more than a decade, placing it in the nascent stage of technical regulation. Compared to the historical experience of international legal regulation in other areas, the path to regulating military AI has just begun. Therefore, the process of establishing customary law through the formulation of new treaties remains a crucial long-term normative pathway that requires persistent international efforts.

---

[126] Gary L. Scott and Craig L. Carr, "Multilateral Treaties and the Formation of Customary International Law", *Denver Journal of International Law & Policy*, Vol. 25, No. 1, 1996, p. 72.
[127] ILC 2018 Report, above note 121, p.136, Commentary to Conclusion 8, para. 3.
[128] ICJ, *North Sea Continental Shelf Cases*, above note 118, para. 73.
[129] The reputational cost for the states will be analysed in part 4.3.
[130] See Michael P. Scharf, "Accelerated Formation of Customary International Law", *ILSA Journal of International & Comparative Law*, Vol. 20, No. 2, 2014, p. 320.

4.2. The Treaty as Evidence: Codification as an Effective Guide for Interpretation

Beyond its direct binding force on party states, the new treaty will serve as a reference for interpreting existing international law, providing an effective guide to its application, and thereby imposing influence on non-party states. To be specific, currently, numerous rules in the field of IHL are codified in treaty texts within the Geneva Conventions and their Additional Protocols. As new weapons emerge, treaty texts require necessary "evolutionary interpretation" based on subsequent agreements and practices to maintain their relevance in new scenarios.[131] In this context, the new convention on military AI will provide a reference for the evolutionary interpretation of existing IHL.

During this process, it should be acknowledged that establishing "subsequent agreement" and "subsequent practice" under Vienna Convention on the Law of Treaties (VCLT) Article 31 imposes a high threshold, requiring that such practice reflect the common understanding of the parties as a whole regarding the interpretation of the treaty.[132] Thus, because the new treaty on military AI may not reflect the agreement of all State Parties to the Geneva Conventions, it cannot constitute a subsequent agreement or practice under VCLT Article 31. However, the practice regarding the new treaty could, in its legal sense, constitute other subsequent practice under VCLT 32, serving as a supplementary means of interpretation for the Geneva Conventions.[133] In its nature, the new treaty could be seen as evidence of states' widespread understanding of their obligations regarding military AI under IHL. In practice, to clarify ambiguous or obscure provisions in the Geneva Conventions, the interpreter could refer to the treaty on military AI. Without distorting the substance of the rules, such supplementary materials may be invoked to clarify the specific implications of existing IHL when applied to the context of military AI. Through this approach, even though some states are not parties to the new treaty, most are signatories to the Geneva Conventions and will thus be influenced by this practice. Should they significantly deviate from the new treaty's provisions in their use of military AI, they may face a greater burden of proof to justify the rationality and legality of such actions.

This interpretative function has particular significance for some broad international rules, such as the Martens Clause. According to the International Law Commission's opinion, the Martens Clause provides that even in cases not covered

---

[131] Such an interpretation practice was early reflected in the WTO U.S. - Shrimp cases. See WTO Appellate Body, *United States - Import Prohibition of Certain Shrimp and Shrimp Products*, WT/DS58/AB/R, 12 October 1998, paras 129-130.

[132] See ILC 2018 Report, above note 121, pp. 20, 28, Commentary to Conclusion 2 para. 9, Commentary of Conclusion 4 para 4.

[133] See *Ibid.*, pp. 20-21, 33-34, Commentary to Conclusion 2 para. 9, Commentary to Conclusion 4 paras. 23-25; *Report of the International Law Commission on the Work of its Sixteenth Session*, UN Doc. A/5809, 11 July 1964, p. 204, para. 13.

by specific international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from the principles of humanity, and from the dictates of public conscience.[134] However, as mentioned in Part 2, there is no universally accepted interpretation of the Martens Clause.[135] Therefore, when international courts are trying to apply such standards to military AI and adjudicating whether a state's application of AWS is consistent with the public conscience, they need to seek references to identify the specific interpretation of that clause. If many states have declared in the new treaty that meaningful human control would be the critical element for complying with the principle of humanity, then such normative expression would affect the courts and state understanding of the Martens Clause, and the new treaty could effectively reduce interpretive discretion.

4.3. Beyond the Law: Imposing Reputational Costs on Non-Party States

A potential treaty on military AI would not only generate legal effects but also impose inescapable reputational costs, thereby promoting treaty compliance even among non-party states. Although the concept of reputational costs has historically remained somewhat ambiguous, international law scholars have consistently identified it as a pivotal mechanism for promoting compliance with international law.[136] As research suggests, in the realm of national security, reputational cost plays a significant role, as in other fields.[137] States that did not shape the current framework of international law face high strategic costs in opting to violate international security law norms perceived as not in their interests.[138] In the field of military AI, based on the simplified cost-benefit theory analysis, if the state pays more reputational cost when it is against the treaty than it gains, then the state is more likely to act consistently with the treaty.[139] So far, the absence of a treaty governing military AI means we may lack empirical means to analyse the reputational costs it generates. Nevertheless, past international law practices offer valuable insights for predicting the reputational costs that new treaties may incur.

Historically, reputational cost was frequently considered when establishing an arms control treaty and has yielded effective results. Not only in the Convention on Cluster Munitions are reputational costs designed as a part of the stigmatizing

---

[134] *Report of the International Law Commission on the Work of its Forty-Sixth Session*, UN Doc. A/49/10, 22 July 1994, p. 317, Commentary to Article 29, para. 3.
[135] Rupert Ticehurst, "The Martens Clause and the Laws of Armed Conflict", *International Review of the Red Cross*, Vol. 37, No. 317, 1997, p. 126.
[136] See Rachel Brewster, "Unpacking the State's Reputation", *Harvard Journal of International Law*, Vol. 50, No. 2, 2009, p. 269.
[137] See Andrew C. Blandford, "Reputational Costs Beyond Treaty Exclusion: International Law Violations as Security Threat Focal Points", *Washington University Global Studies Law Review*, Vol. 10, No. 4, 2011.
[138] *Ibid.*, p. 726.
[139] See Andrew T. Guzman, *How International Law Works: A Rational Choice Theory*, Oxford University Press, Oxford, 2008, pp. 74-75.

mechanism,[140] but also in the Mine Ban Treaty, reputational costs substantially facilitated the ban on mines.[141] Specifically, the Mine Ban Treaty gives the signatory states a solid moral and political basis for criticizing other states' ownership or use of these weapons, increasing the cost for opposing states of using them.[142] Hence, when the Mine Ban Treaty was adopted, it resulted in an abrupt change of the UK's position on mines. The humanitarian concerns even affected its domestic politics, which pushed the UK to join the treaty, and resulted in the United States being isolated from its NATO allies.[143] Eventually, the Mine Ban Treaty successfully forced countries like China, India, and the US to make a substantial change in their policy.[144]

Regarding the potential treaty on military AI, we should admit that military AI could provide much more military utility than cluster munitions and mines, which could affect the efficacy of reputational costs. However, this article advocates for the regulation of military AI, rather than a complete ban akin to those on cluster munitions and landmines. In this sense, if a state can regulate certain aspects of military AI without compromising its core military interests while simultaneously gaining additional reputational benefits, it can take such action based on its rational choice. Therefore, the mechanism by which reputational costs influence the regulation of military AI does not directly ensure that states fully control such technologies. Instead, it indirectly prompts states to make choices aligned with international expectations by increasing or decreasing these costs. A potential treaty on military AI serves as an amplifier in this process.

5. Conclusion

The military application of artificial intelligence challenges the international legal system, splits states, as well as scholars, between advocating for a new treaty and relying on existing international law. The opponents of establishing a new treaty argue that existing international law is sufficient, while the proponents contend that the new problems posed by military AI necessitate a new treaty. This article examines the potential problems posed by military AI and argues that the moral dilemmas, accountability gaps, and compliance deficits would necessitate the establishment of a legally binding instrument on a *de jure* basis to diminish legal uncertainty.

However, from a *de facto* perspective, establishing a new treaty in the near future faces too many obstacles to be feasible, making the establishment of a treaty

---

[140] ICRC, Convention on Cluster Munitions, Geneva, 2008, Foreword, p. 6.
[141] Arms Control Association, "The Ottawa Convention at a Glance", 2024, available at: https://www.armscontrol.org/factsheets/ottawa-convention-glance.
[142] Kenneth R. Rutherford, "The Evolving Arms Control Agenda: Implications of the Role of NGOs in Banning Antipersonnel Landmines", *World Politics*, Vol. 53, No. 1, 2000, pp. 109-110.
[143] *Ibid.*, p. 100-101.
[144] *Ibid.*, n. 131.

that can directly regulate major powers almost improbable. Under the CCW process, there has been no mandate to negotiate a treaty, and the consensus requirements for adopting one give the major military powers *de facto* veto power, making it almost impossible to negotiate a new treaty. Under the alternative process, even though a treaty can be established like the Convention on Cluster Munitions, it is doubtful to what extent such a treaty *per se* would effectively regulate its opposing states. From a realist perspective, since military AI is a strategic imperative for major military powers, expecting these countries to restrict technologies that would significantly increase their battlefield advantages is irrational. Under the landscape of contestation over international rule-making power, the universal scepticism among Western states makes the establishment of such a treaty even harder. At the macro level, the intensification of geopolitical rivalry and an unfavourable political environment also serve as negative factors affecting treaty-making. All those elements contribute to the international regulation of military AI devolving into a "majority follows minority" process.

Despite obstacles mentioned, the article still holds a cautiously optimistic note regarding the establishment of a new treaty on military AI. It argues that the *de facto* improbability does not nullify the *de jure* necessity. Instead, the very pursuit of a new treaty initiates a dynamic legal process. In the long term, such a legal process would impose legal and non-legal force on the party and non-party states through three approaches, all of which would facilitate the international regulation of military AI. Firstly, the new treaty could serve as an incubator for CIL. The treaty-making process will promote the generation and crystallization of CIL through enhancing state practice and *opinio juris*. Secondly, the treaty text could serve as evidence of existing international law during legal interpretation, thereby diminishing the possibility of arbitrary interpretation of the existing law by treaty opponents. Finally, beyond the legal effects that it could have on non-party states, it could also create reputational costs that make states think twice before openly acting inconsistently with it.

The Concept of Lethality in the Autonomous Weapon Systems Debate and Its Implications

Mei Ching Liu[*]

ABSTRACT

This article explores the concept of lethality in the autonomous weapon systems (AWS) debate by examining the arguments for and against the inclusion of this concept within the definition of these systems. The arguments both for and against such inclusion carry significant implications for international security, as the definition of AWS will directly shape the scope of any future regulatory framework. As there appears to be a provisional agreement to include the concept, this article examines the proposed definitions and evaluates the implications of their adoption—ranging from the architecture of a potential regulatory framework to the risk of conflating international humanitarian law and international human rights law.

Keywords: Autonomous weapon systems, Lethal autonomous weapon systems, Artificial intelligence, Lethality, Non-lethal weapon

I. Introduction

For almost a decade, multilateral dialogue surrounding the use of artificial intelligence (AI) in the military domain has focused, almost entirely, on autonomous weapon systems (AWS). These systems are commonly understood as systems that, once activated, can autonomously select and engage targets without further human intervention. [1] While AWS may or may not incorporate AI, technological advancements, particularly in AI and robotics, have raised humanitarian, legal and security concerns surrounding their development and use thereby successfully capturing public and political attention.

Multilateral dialogue on AWS is led by the United Nations (UN) group of governmental experts (GGE) related to emerging technologies in the area of lethal autonomous weapons systems (LAWS), under the auspices of the Convention on Certain Conventional Weapons (CCW)—a key international humanitarian law

[*] Mei Ching Liu is an Associate Research Fellow with the Military Transformations Programme at the S. Rajaratnam School of International Studies (RSIS), Nanyang Technological University (NTU), Singapore. All views presented in this article are the author's own and do not represent the official position of RSIS, NTU. E-mail: ismeiching.liu@ntu.edu.sg.

[1] International Committee of the Red Cross (ICRC), *ICRC Position and Background Paper on Autonomous Weapon Systems*, Geneva, 12 May 2021, p. 2. AWS encompass various existing military technologies. These include: (i) air defence systems used on warships or at military bases that can be used to strike incoming missiles, rockets or mortars; (ii) active protection systems on armoured vehicles designed to strike incoming munitions; (iii) certain guided munitions; and (iv) mines which have been described as less sophisticated AWS. See *ibid.*, p. 5; Stockholm International Peace Research Institute (SIPRI), *Mapping the Development of Autonomy in Weapon Systems*, Solna, November 2017.

(IHL) instrument which seeks to govern the use of certain types of weapons that are deemed to cause unnecessary or unjustifiable suffering, or to affect civilians indiscriminately.[2] The GGE was established by the High Contracting Parties to the CCW in 2016.[3]

It has been a decade since the GGE was established, and the group is still unable to reach a consensus on the regulation of these weapon systems.[4] A key point of disagreement is whether the term "lethal" should be used to describe AWS, meaning whether these systems should be described as "lethal" AWS (LAWS) or simply AWS. Arguments offered by States generally fall along two fault lines. One group of States insists on incorporating the term "lethal" into the definition,[5] citing the GGE's mandate concerning emerging technologies in the area of LAWS. Conversely, another group opposes this, arguing that the concept of lethality lacks a basis in IHL and lethality is an effect of a weapon rather than a design feature.

The GGE is currently mandated to "further consider and formulate, by consensus, a set of elements of an instrument…and other possible measures to address emerging technologies in the area of lethal autonomous weapon systems".[6] As the GGE's current mandate ends in 2026 with growing calls from civil society and the UN Secretary-General for a legally binding instrument on these weapon systems,[7] this article assesses the debate surrounding the issue of lethality at the GGE. In doing so, Section 2 of the article maps the state of the lethality debate within

---

[2] 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 (CCW), 1342 UNTS 137, 10 October 1980 (entered into force 2 December 1983), Preamble.

[3] Final Document of the Fifth Review Conference, UN Doc. CCW/CONF.V/10, 23 December 2016, p. 9.

[4] Since its establishment, the GGE has affirmed, in 2017, that IHL continues to apply fully to all weapons systems. In 2019, the GGE developed a set of guiding principles to address the challenges posed by these systems, particularly those related to accountability for their development and deployment. In 2023, the GGE affirms, among others, the applicability of IHL to the potential development and use of LAWS, and control with regard to weapon systems, including LAWS. See Report of the 2017 Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS), UN Doc. CCW/GGE.1/2017/3, 22 December 2017; Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, UN Doc. CCW/GGE.1/2019/3, 25 September 2019; Report of the 2023 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, UN Doc. CCW/GGE.1/2023/2, 24 May 2023.

[5] It is noted that States generally refer to the term "definition" as a "working definition" or "working characterisation". This article uses the term "definition" instead of "working definition" for ease of reference, and to align with the language used when the GGE was first established. See Report of the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) (Report of the 2016 Informal Meeting), UN Doc. CCW/CONF.V/2, 10 June 2016, paras. 35-42.

[6] Meeting of the High Contracting Parties to the CCW—Final Report, UN Doc. CCW/MSP/2023/7, 23 December 2023, para. 20.

[7] "Secretary-General's Remarks at the Launch of the Policy Brief on a New Agenda for Peace", 20 July 2023, available at: https://www.un.org/sg/en/content/sg/statements/2023-07-20/secretary-generals-remarks-the-launch-of-the-policy-brief-new-agenda-for-peace-bilingual-delivered-follows-scroll-down-for-all-english-and-all-french (all internet references were accessed on 13 July 2026).

the GGE under its current mandate (2024–2026) and tracks the changes in position. Thereafter, Section 3 examines whether the attribute of lethality is required by analysing the documents pertaining to the establishment of the GGE. This examination is conducted to evaluate the claim that lethality is central to the GGE's mandate. Subsequently, Sections 4 and 5 analyse the arguments and counterarguments related to the inclusion of this concept within the definition before attempting to define the concept in Section 6. Section 7 concludes with a set of recommendations for policymakers.

This article demonstrates that while the decision to include or exclude the concept of lethality within the definition of AWS is purely policy-related, such a decision carries both policy and legal implications for the broader weapon regulatory framework. This article does not aim to derail the discussions currently undergoing within the GGE but rather hopes that the analysis provided herein may assist in delineating policy and legal arguments so that the object and purpose of the CCW, under which the GGE operates, may be achieved.

2. The Lethality Debate at the GGE (2024 – 2026)

This section provides an overview of the current debate surrounding the concept of lethality under the GGE's current mandate. It examines the arguments advanced by States both for and against including the concept of lethality within the definition of AWS. The analysis provided in this section finds that these State positions are, at times, inconsistent.

To support the GGE in fulfilling its mandate, the GGE Chairperson introduced a document in 2024, which is referred to as the "rolling text". It lays out a set of elements that could be incorporated into a future final instrument regulating AWS. These elements include, among others, the definition of AWS and the types of AWS that should be prohibited and regulated.

The Chairperson circulated the first version of the rolling text on 26 July 2024,[8] the second version on 8 November 2024,[9] the third version on 12 May 2025,[10]

[8] "GGE on LAWS, Rolling Text, Status Date: 26 July 2024" (Rolling Text – 26 July 2024), available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_ Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons _Systems_(2024)/Rolling_text_draft.pdf.

[9] "GGE on LAWS, Rolling Text, Status Date: 8 November 2024" (Rolling Text - 8 November 2024), available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional _Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons _Systems_(2024)/Revised_rolling_text_as_of_8_November_2024_final.pdf.

[10] "GGE on LAWS, Rolling Text, Status Date: 12 May 2025" (Rolling Text - 12 May 2025), available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_ Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_ Systems_(2025)/CCW_GGE_LAWS_-_Revised_rolling_text_as_of_12_May_2025.pdf.

and the fourth version on 18 December 2025.[11] The evolution of the definition can be seen in Table 1 below:[12]

Table 1: Evolution of the Definition

Rolling Text – 26 July 2024

Without prejudging any other options for measures, a lethal autonomous weapon system can be characterised as a weapon system that, after activation, can select and engage a target without further intervention by a human operator.

Rolling Text – 8 November 2024

1. A lethal autonomous weapon system can be characterised as an integrated combination of one or more weapons and technological components that enable the system to identify and/or select, and engage a target, without intervention by a human user in the execution of these tasks.

Rolling Text – 12 May 2025

1. Within the scope of the application of the CCW, ~~A~~ a lethal autonomous weapon system can be characterised as an integrated combination of one or more weapons and technological components, that can select~~,~~ and engage a target, without intervention by a human user in the execution of these tasks.

2. For the purposes of these elements of an instrument, lethal refers to the capacity to cause death to a person. The fact that a LAWS can be used in circumstances that do not result in death, such as to damage or destroy objects or to cause injury, does not exclude it from this characterisation.

Rolling Text – 18 December 2025

1. [Within the scope of the application of the CCW and for the purpose of these elements of an instrument,] a lethal autonomous weapon system can be characterised

[11] "GGE on LAWS, Rolling Text, Status Date: 18 December 2025" (Rolling Text - 18 December 2025), available at: https://docs-library.unoda.org/Convention_on_Certain_ Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_ Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_ 2025.pdf.

[12] The definitions provided in the rolling text have been standardised to British English. At the time of the editorial process, the fifth version of the rolling text has been published. See "GGE on LAWS, Rolling Text, Status Date: 05 June 2026", available 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.

as a functionally integrated combination of one or more weapons and technological components, that can identify, select, and engage a target, without intervention by a human operator in the execution of these tasks.

…

B. The fact that a LAWS can be used in a way that does not result in loss of life, such as to damage or destroy objects or to cause injury, does not exclude it from this characterisation.

During the GGE's second meeting for 2024 held in August, Switzerland opposed the inclusion of the term "lethal" in the definition.[13] It noted that lethality is an effect of using a weapon system rather than a characteristic of such a system and that "[m]any IHL challenges are much broader than lethality",[14] and therefore, the group must not limit its considerations solely to this concept. Ireland clarified that IHL also protects against superfluous injury and unnecessary suffering, as well as providing protection for objects.[15] Therefore, it considered the inclusion of "lethal" as a qualifier in the definition to be inappropriate. Other states, such as Austria,[16] Brazil,[17] Canada,[18] Costa Rica,[19] Guatemala,[20] New Zealand,[21] Norway,[22] Pakistan,[23] Panama,[24] and Peru,[25] also opposed including the term "lethal".

In contrast, Israel argued that the GGE must focus on "lethal" AWS and "not just any autonomous weapons system in accordance with [the group's] current

[13] "Switzerland", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=2539.
[14] *Ibid.*
[15] "Ireland", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=3876.
[16] "Austria", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=4557.
[17] "Brazil", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=4899.
[18] "Canada", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/FA0FFCD3-4DA7-4ADF-96F9-03AA13398337/10h19/en?position=1245.
[19] "Costa Rica", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=3135.
[20] "Guatemala", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=6043.
[21] "New Zealand", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=7240.
[22] "Norway", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=9569.
[23] "Pakistan", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=4033.
[24] "Panama", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=7350.
[25] "Peru", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=4367.

mandate".[26] It emphasised that the inclusion of non-lethal force could lead to broadening the GGE's debate to encompass weapon systems that fall outside the purview of the GGE's mandate. Furthermore, it warned that by deviating from the issue of lethality, the GGE risked "losing sight of compliance with the key aspects of using weapons under IHL".[27] India supported Israel's remarks.[28] China similarly supported the inclusion of "lethal" in the definition. It noted that a "lethal weapon system refers to a weapon system that is designed for and has the function of killing and maiming, and it is indeed capable of causing such killing and maiming effect when it carries out its mission."[29]

During the meeting, the GGE considered four alternative formulations of the definition but failed to reach a consensus.[30] In November 2024, the Chairperson circulated an amended definition in the second version of the rolling text. This was intended for the GGE's consideration prior to its March 2025 meeting.

---

[26] "Israel" (Israel's Statement at the GGE), available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=5139.

[27] *Ibid.*

[28] "India" (India's Statement at the GGE), available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/7D069CCD-CAD1-4EF2-9027-0DD859BFB8E3/15h13/en?position=7538.

[29] "China" (China's Statement at the GGE), available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/B2BE5F31-6F30-4EF4-BEF1-5D648F300165/15h16/en?position=5004.

[30] Reaching Critical Will, *Civil society perspectives on the Group of Governmental Experts of the Convention on Certain Conventional Weapons on Lethal Autonomous Weapon Systems 26–30 August 2024*, Geneva, 5 September 2024, pp. 11-28. The first formulation of the definitions can be seen in Table 1 above, under "Rolling Text – 26 July 2024". The subsequent formulations were presented by the GGE Chairperson on 27, 28 and 30 August 2024 respectively. The second formulation can be seen here: "Rolling Text, Status Date: 27 August 2024", available at: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2024/gge/documents/rolling-text-characterisation.pdf. The third formulation can be seen here: "Rolling Text, Status Date: 28 August 2024", available at: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2024/gge/documents/rolling_text_characterisation_3.pdf. The final formulation can be seen here: "Rolling Text, Status Date: 30 August 2024", available at: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2024/gge/documents/rolling_text_characterisation_4.pdf.

During the group's meeting in March, the issue of lethality was extensively debated. States, such as China,[31] India,[32] Russia,[33] Singapore,[34] and the United States,[35] insisted on retaining the term "lethal". They considered it a core component of these weapon systems and central to the GGE's mandate. In contrast, other States, including Austria,[36] Brazil,[37] Pakistan,[38] and Switzerland,[39] opposed it.

Consequently, the group considered an alternative iteration of the definition, which also included a specific definition of the term "lethal".[40] Once again, the group failed to reach a consensus. Following this, and ahead of the informal consultations on these weapon systems in New York in May 2025,[41] the GGE Chairperson

---

[31] "China", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=6119.

[32] "India", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=8411.

[33] "Russia", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=9303.

[34] "Singapore", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=7716. Singapore has also suggested that lethal weapons mean weapons that are designed to kill or intended to kill. See "Singapore", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/735DA0AD-9CD1-4F91-A5EB-743C519D65ED/10h31/en?position=3044.

[35] "United States", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=3874.

[36] "Austria", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/2008D96D-0896-49F4-B2DD-FB23082C8DA0/15h11/en?position=2403.

[37] "Brazil", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=2618.

[38] "Pakistan", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=2945.

[39] "Switzerland", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06?position=4999.

[40] Reaching Critical Will, *Civil Society Perspectives on the Group of Governmental Experts of the Convention on Certain Conventional Weapons on Lethal Autonomous Weapon Systems 3–7 March 2025*, Geneva, 11 March 2025, pp. 4-8. The alternative iteration of the definition that was debated by the GGE is as follows:

> "1. Within the scope of the application of the CCW, a lethal autonomous weapon system can be characterised as an integrated combination of one or more weapons and technological components that can identify and/or select, and engage a target, without intervention by a human user in the execution of these tasks.
>
> 2. For purposes of these elements of an instrument, lethal means weapons with the capacity to kill or injure one or more persons, which includes weapons that can be used to damage or destroy objects."

[41] In accordance with the United Nations General Assembly (UNGA) Resolution 79/62, a two-day informal consultation was held in New York to consider the UN Secretary-General's report on LAWS "in full complementarity with and in a manner that supports the fulfilment of the mandate" of the GGE. See "Informal Consultations on Lethal Autonomous Weapons Systems", available at: https://unodaweb-meetings. unoda.org/public/2025-03/20250324%20AM.pdf; UNGA Res. 79/62, 2 December 2024. For the UN Secretary-General's Report on LAWS, see Lethal Autonomous Weapons Systems – Report of the Secretary-General, UN Doc. A/79/88, 1 July 2024.

circulated an amended definition within the third iteration of the rolling text. This was intended for the GGE's consideration ahead of its second meeting of 2025.

During the GGE's second meeting in September 2025, some of the States opposing the inclusion of "lethal" signalled their willingness to compromise.[42] This shift was likely due to the addition of a definition for the term "lethal", alongside a clarifying sentence which followed immediately after the definition, as set out in the third version of the rolling text. The paragraph, when read as a whole, would encompass not only the death of persons but also injury, as well as the damage or destruction of objects.

This wider scope addressed the concerns of States that opposed the inclusion of the term "lethal" by moving the focus away from human fatalities to a more comprehensive one that includes a broader range of effects of using these weapon systems. For these States, the deletion of the clarifying sentence—namely, that "a LAWS can be used in circumstances that do not result in death, such as to damage or destroy objects or to cause injury, does not exclude it from this characterisation"—was unacceptable.[43] In essence, the underlying position of these States did not shift. Rather, the inclusion of the term "lethal" can be viewed as a strategic diplomatic move aimed at achieving consensus. Their concerns—specifically that IHL also protects objects and guards against human injury—were effectively addressed by the clarifying sentence.

For those States insisting on including the term "lethal" in the definition of AWS, their contention is that this term is central to the GGE's mandate and serves to limit the group's scope.[44] However, the paragraph containing the definition of the term "lethal" and its clarifying sentence was modelled on a suggestion from these States.[45] If the primary aim of including the term "lethal" was indeed to limit the

---

[42] Mei Ching Liu, "International Regulation of Lethal Autonomous Weapon Systems: A Long Road Ahead", *IDSS Paper*, 1 October 2025, available at: https://rsis.edu.sg/rsis-publication/idss/ip25095-international-regulation-of-lethal-autonomous-weapon-systems-a-long-road-ahead/.

[43] For example, see "Spain", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/F1F4A6EF-06DC-47DB-B361-30B89BAF4CC4/10h14/en?position =4819.

[44] For example, Israel's Statement at the GGE, above note 26; India's Statement at the GGE, above note 28; China's Statement at the GGE, above note 29; "Working Paper of the People's Republic of China on Lethal Autonomous Weapons Systems (July 2022)", p. 2, available at: https://documents.unoda.org/wp-content/uploads/2022/07/Working-Paper-of-the-Peoples-Republic-of-China-on-Lethal-Autonomous-Weapons-Systems%EF%BC %88English%EF%BC%89.pdf.

[45] Specifically, the United States. It proposed it during the GGE's meeting on 3 March 2025. The relevant parts of the United States' proposal are as follows: "I think we would suggest that one way to get at this would be to include after paragraph 1, 1*bis* that could read 'Lethal force is force that is capable of resulting in death. The fact that LAWS can be used in circumstances that do not result in death such as to destroy objects or to cause injury does not exclude it from its characterisation.' And that we think would be a way to get at the concern that we think animates the focus on lethality. This would acknowledge that these types of weapon systems can also be used in non-lethal operations and that they wouldn't be excluded for that reason." See "United States", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE4/10h06/en?position=3874.

scope of the GGE's discussions, the definition in the third version of the rolling text—along with its clarifying sentence—would fail to achieve that restriction. Therefore, it ought to have received an overwhelming rejection from States which insisted on the term's inclusion.

Surprisingly, this did not happen. As noted in the GGE Chairperson's summary of the September 2025 meeting, "there was general support" for the definition of lethality in the third version of the rolling text, which marked a break from the previous disagreements,[46] but the reason for the shift in States' positions was unclear as these States have not explained their reasons for such a shift.

Nevertheless, in the March 2026 meeting, some States supportive of including the term "lethal" in the definition sought further concessions from those that opposed it. For example, there was a proposal specifically aiming to restrict the effects of such weapon systems to exclude objects.[47] States, including those which opposed the inclusion of the term "lethal" in the definition, rejected this suggestion.[48] Another proposal concerned the removal of the phrase "does not result in loss of life" in the clarifying sentence; this was also rejected by some States.[49]

---

[46] Chair's Summary – Second 2025 Session of the GGE on LAWS, UN Doc. CCW/GGE.1/2025/WP.9, 20 October 2025, para. 12. During the March 2025 meeting, States such as Japan and the Republic of Korea, which had insisted on the inclusion of the term "lethal", had expressed support for the United States' proposal. See "Japan" (Japan's Statement at March 2025 Meeting), available at: https://conf.unog.ch/ digitalrecordings/en/clients/61.0500/meetings/5A654DCE-4B95-4676-9685-CFF7439A8DE 4/10h06/en?position=8245; "Republic of Korea", available at: https://conf.unog.ch/ digitalrecordings/en/clients/61.0500/meetings/735DA0AD-9CD1-4F91-A5EB-743C519D65E D/10h31/en?position=5838. This support also persisted in the March 2026 meeting. For example, Japan stated that "Mr. Chair, point number 2, quite briefly regarding lethality, while my Delegation does not object to including nonfatal injury and material damage within the scope, but we would like to clarify the focus on this discussion on the lethality as included in the very mandate of this GGE." See "Japan" (Japan's Statement at March 2026 Meeting), available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/ meetings/83ECCAE6-057E-4EA5-840C-351247F34438/15h03/en?position=3984.

[47] For example, during the session on 5 March 2026, Russia proposed to remove the term "target" and replace it with "manpower", "human being" or "personnel". See "Russia", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/86B3368 4-84E0-498A-B6B0-327A27D50405/10h13/en?position=2220.

[48] For example, Brazil, Switzerland and the United Kingdom. See "Brazil", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498 A-B6B0-327A27D50405/10h13/en?position=4116; "Switzerland", available at: https:// conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498A-B6B0-327A27D50405/10h13/en?position=4570; "United Kingdom", available at: https://conf. unog.ch/digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498A-B6B0-327A 27D50405/10h13/en?position=3499.

[49] For example, Austria and Norway. See "Austria", available at: https://conf. unog.ch/digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498A-B6B0-327A27D50405/10h13/en?position=7878; "Austria", available at, https://conf.unog.ch/ digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498A-B6B0-327A27D50405/10h13/en?position=8795; "Norway", available at: https://conf.unog.ch/ digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498A-B6B0-327A27D5040 5/10h13/en?position=6923.

3. The Lethality Debate Before the Establishment of the GGE (2013 – 2016)

As discussed in Section 2, one key argument advanced by States which insisted on including the term "lethal" within the definition is that the concept of lethality is central to the GGE's mandate. To evaluate this claim, this section examines the history of the GGE's establishment.

Issues surrounding AWS were first raised within the international human rights law (IHRL) framework in a 2013 report by the UN Special Rapporteur Christof Heyns, which was submitted to the UN Human Rights Council.[50] In this report, Heyns referred to these weapon systems as "lethal autonomous robotics" based on his observations regarding the implications for the protection of life under the IHRL framework,[51] stemming from targeted killings conducted with remotely piloted armed drones.[52] The systems referred to in the report were initially developed for surveillance and were considered non-lethal until their use expanded to include armed strikes.[53] As these systems were primarily used in warfare,[54] consensus emerged among States that issues surrounding AWS would be better addressed within the IHL framework rather than the IHRL framework, as IHL is the legal regime specifically crafted to regulate the conduct of warfare. As a result, multilateral discussions related to AWS were shifted to another forum, i.e. the CCW.[55]

When concerns regarding AWS were introduced during the 2013 meeting of the High Contracting Parties to the CCW (held from 14 to 15 November 2013), these systems were described using 13 different labels.[56] Table 2 below identifies the terms

---

[50] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, UN Doc. A/HRC/23/47, 9 April 2013. Heyns also submitted a report to the UN General Assembly in 2013, see Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions (Heyns' 2013 Report to UNGA), UN Doc. A/68/382, 13 September 2013.

[51] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, above note 50, paras. 26-36.

[52] *Ibid.,* paras. 44-52, Heyns' 2013 Report to UNGA, above note 50, paras. 12-25.

[53] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, above note 50, paras. 29-30; Heyns' 2013 Report to UNGA, above note 50, paras. 12-25.

[54] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, above note 50, paras. 50-81. Heyns also outlined the use of armed drones outside of warfare, see Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, above note 50, paras. 82-85.

[55] "Timeline of LAWS in the CCW", available at: https://disarmament.unoda.org/timeline-of-laws-in-the-ccw/.

[56] Based on an analysis of the documents and statements presented during the 2013 meeting. There were no audio recordings. Given the current conceptual understanding of AWS, labels exclusively associated with drones, such as "armed drones" and "weaponised drones", were not included. For the documents and statements presented during the 2013 meeting, see "Convention on Certain Conventional Weapons — Meeting of High Contracting Parties (2013)", available at: https://meetings.unoda.org/meeting/29457/documents. For availability of audio recordings, see "Digital Recordings Portal", available at: https://conf.unog.ch/digitalrecordings/en/search?date%5Bmin%5D =2013-11-14&date%5Bmax%5D=2013-11-15&keywords=Meeting+of+the+High+Contracting+Parties+to+the+CCW&audience=&organization=All.

used to describe these weapon systems and the parties that used these terms during the meeting

Table 2: Terms that were used to Describe AWS in 2013 CCW Meeting

No.	Terms	States/Non-governmental Organisations
1	Autonomous weapons	International Committee of the Red Cross (ICRC),[57] United States[58]
2	AWS	UN Secretary-General, [59] Austria, [60] Article 36, [61] Ghana, [62] Canada, [63] IKV Pax Christi,[64] China,[65] United States[66]

[57] "Statement of the ICRC to the 2013 Meeting of States Parties to the Convention on Certain Conventional Weapons", p. 1, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/ICRC_MSP%2B2013_statement.pdf

[58] "U.S. Delegation Opening Statement", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/USA_MSP_GenStatement_2013.pdf. This description was used as a subheading in the statement.

[59] "The Secretary-General – Message to the Conferences of the High Contracting Parties to the Convention on Certain Conventional Weapons and Its Protocols", available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_ Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/DGS_MSP_ GenStatement_2013.pdf.

[60] "Statement by Austria", p. 3, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Austria_MSP%2B2013_statement.pdf.

[61] Meeting of the High Contracting Parties to the CCW – Summary Record of the 2nd Meeting (Summary Record of the 2nd Meeting), UN Doc. CCW/MSP/2013/SR.2, 4 December 2013, para. 28; "Statement by Article 36 to the Meeting of States Parties to the CCW", p. 1, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_ Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/NGO article36_MSP_GenStatement_2013.pdf.

[62] "Statement by the Delegation of Ghana", p. 1, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Ghana_MSP%2B2013_statement.pdf.

[63] Summary Record of the 2nd Meeting, above note 61, para. 5.

[64] *Ibid.*, para. 33.

[65] Réunion des Hautes Parties contractantes à la Convention sur l'interdiction ou la limitation de l'emploi de certaines armes classiques qui peuvent être considérées comme produisant des effets traumatiques excessifs ou comme frappant sans discrimination - Compte rendu analytique de la 1re séance (Summary Record of the 1st Meeting), UN Doc. CCW/MSP/2013/SR.1, 19 November 2013, para. 24.

[66] *Ibid.*, para. 26.

3	Fully autonomous weapons	Republic of Lithuania, [67] IKV Pax Christi, [68] Human Rights Watch, [69] European Union,[70] ICRC[71]
4	Fully AWS	International Committee for Robot Arms Control,[72] Republic of Korea[73]
5	Killer robots	Ukraine[74]
6	Lethal autonomous robots	Egypt,[75] ICRC,[76] Human Rights Watch[77]
7	Lethal autonomous robotics	Germany,[78] Japan,[79] Pakistan,[80] IKV Pax Christi[81]

[67] "Statement by the Republic of Lithuania", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Lithuania_MSP%2B2013_statement.pdf.
[68] Summary Record of the 2nd Meeting, above note 61, para. 33.
[69] *Ibid.*, paras 21-25; "Statement by Steve Goose – Human Rights Watch", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/NGOHRW_MSP_GenStatement_2013.pdf.
[70] "EU Statement at the Meeting of the High Contracting Parties to the Convention on Certain Conventional Weapons", p. 4, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/EU_MSP%2B2013_statement.pdf.
[71] "Statement of the ICRC to the 2013 Meeting of States Parties to the Convention on Certain Conventional Weapons", above note 57, p. 1.
[72] Summary Record of the 2nd Meeting, above note 61, paras. 29-32.
[73] "Meeting of the High Contracting Parties to the CCW – General Exchange of Views", para. 5, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/RoK_MSP%2B2013_statement.pdf.
[74] "Statement by Mr Oleksandr Aleksandrovich – Head of Delegation of Ukraine", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Ukraine_MSP%2B2013_statement.pdf.
[75] "Statement of Egypt", available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Egypt_MSP_LAWS_2013.pdf.
[76] Summary Record of the 2nd Meeting, above note 61, para. 18. The term "lethal autonomous robots" was used as an alternative for autonomous weapons.
[77] *Ibid.*, para. 25. It is stated that "fully autonomous weapons, also known as lethal autonomous robots".
[78] "General Statement by Germany", p. 3, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Germany_MSP%2B2013_statement.pdf.
[79] "Statement by H.E. Ambassador Toshio Sano – Permanent Representative of Japan to the Conference on Disarmament", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Japan_Msp_GenStatement_2013.pdf.
[80] "Statement by Ambassador Zamir Akram – Permanent Representative", p. 3, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Pakistan_MSP%2B2013_statement.pdf.
[81] Summary Record of the 2nd Meeting, above note 61, para. 33.

8	Lethal autonomous weapons	Italy,[82] Holy See,[83] Türkiye,[84] Mexico[85]
9	LAWS	UN Inter-Agency Coordination Group on Mine Action, [86] UN Mine Action Service, [87] Ireland, [88] France, [89] Brazil, [90] Switzerland, [91] European Union, [92] Belarus,[93] Lithuania[94]

[82] "Statement by Ambassador Vincio Mati – Permanent Representative of Italy to the Conference on Disarmament", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Italy_MSP%2B2013_statement.pdf. Italy also described lethal autonomous weapons as "the so-called killer robots".

[83] "Statement by H. E. Archbishop Silvano M. Tomasi", p. 1, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_ Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/HolySee_MSP%2B2013_ statement.pdf. Throughout the statement, varied terminologies were used, including robotic weapon systems and weapon drones.

[84] "Statement by Turkey", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Turkey_MSP%2B2013_statement.pdf.

[85] "Intervención de México", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Mexico_MSP%2B2013_statement.pdf.

[86] "United Nations Inter-Agency Coordination Group on Mine Action", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/IACG_MSP_GenStatement_2013.pdf.

[87] "United Nations Inter-Agency Coordination Group on Mine Action – Delivered by UNMAS", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/UNMAS_MSP_Statement_2013.pdf; Summary Record of the 2nd Meeting, above note 61, para. 15.

[88] "Statement by Ireland", p. 3, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Ireland_MSP_GenStatement_2013.pdf.

[89] "Convention sur Certaines Armes Classiques – Intervention Générale", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/France_MSP%2B2013_statement.pdf.

[90] "Statement by Ambassador Antonio Guerreiro – Permanent Representative of Brazil to the Conference on Disarmament", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Brazil_MSP%2B2013_statement.pdf.https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Brazil_MSP%2B2013_statement.pdf

[91] "Switzerland – Informal Copy for Interpreters", available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Switzerland_MSP_GenStatement_2013.pdf; Summary Record of the 2nd Meeting, above note 61, para. 11.

[92] "EU Statement at the Meeting of the High Contracting Parties to the Convention on Certain Conventional Weapons", above note 70, p. 4.

[93] Summary Record of the 2nd Meeting, above note 61, para. 7.

[94] "Statement by the Republic of Lithuania", above note 67, p. 2.

10	Lethal fully autonomous robots	Australia[95]
11	Lethal fully AWS	United States[96]
12	Robots	Cuba[97]
13	Robotic weapons systems	Human Rights Watch,[98] Holy See[99]

As seen in Table 2, the majority of States and non-governmental parties used one or more terminologies to describe AWS.

During the same meeting, the Chairperson proposed to hold an informal meeting of experts in 2014, but this proposal itself was referred to using a range of titles. In the 2013 meeting's final report, it was called "emerging technologies in the area of lethal autonomous weapons systems"[100]—this is also the name of the GGE. However, when listed as a provisional agenda item for the 2014 meeting of the CCW High Contracting Parties, it was simply referred to as "LAWS".[101] Additionally, the 2013 meeting's summary record further detailed other descriptions used by various parties when addressing the Chairperson's proposal:

- "Canada supported the proposal to hold an informal meeting of experts to discuss matters related to emerging lethal weapons technologies, including autonomous weapons systems."[102]
- "As a member of the Campaign to Stop Killer Robots, Article 36 welcomed the proposal that the Meeting of the High Contracting Parties should adopt a mandate to discuss autonomous weapons systems."[103]
- "ICRAC welcomed the proposal to convene expert meetings in 2014 to examine the challenges posed by fully autonomous weapons systems."[104]

No audio recording or written document is presently available to ascertain the precise title of the Chairperson's proposal.

---

[95] "Statement by Mr Ian McConville, Chargé, Australian Permanent to the United Nations in Geneva", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Australia_MSP%2B2013_statement.pdf.

[96] "U.S. Delegation Opening Statement", above note 58, p. 2.

[97] "Intervencion de Nadiezda Arredondo, Miembro de la Delegación de Cuba", p. 1, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Meeting_of_High_Contracting_Parties_(2013)/Cuba_MSP_Statement_2013.pdf.

[98] "Statement by Steve Goose – Human Rights Watch", above note 69, p. 2.

[99] "Statement by H. E. Archbishop Silvano M. Tomasi", above note 83, p. 3.

[100] Meeting of the High Contracting Parties to the CCW – Final Report (2013 CCW Final Report), UN Doc. CCW/MSP/2013/10, 16 December 2013, paras. 18 and 32.

[101] *Ibid.*, p. 6.

[102] Summary Record of the 2nd Meeting, above note 61, para. 5.

[103] *Ibid.*, para. 28.

[104] *Ibid.*, para. 32.

As a result of the 2013 discussion, the first informal meeting of experts was convened in 2014 under the CCW. At the 2014 informal meeting of experts, the terminology used in the meeting report regarding the mandate was not consistent. First, it referred to the mandate as "emerging technologies in the area of lethal autonomous weapons systems", citing the 2013 meeting's final report.[105] Yet, the title of the report did not include the "emerging technologies" aspect, and within the same report, it was stated that "delegations were reminded that the mandate was about emerging technologies"; there was no mention of LAWS and/or AWS.[106] The programme included as an annex to the report only used the term "LAWS".[107] Throughout this meeting, these weapon systems continued to be described with a range of different terminologies by various States and non-governmental parties.[108]

The use of different terminologies persisted into the 2015 meeting.[109] In 2016, the informal meeting of experts considered whether the attribute of lethality was a required aspect of the mandate, based on discussion captured in the meeting report:

> "41. The question was raised as to whether the attribute of lethality was required and that instead it would be more constructive to focus on the use of force. There was a proposal to understand LAWS more inclusively, covering also means and methods of warfare that do not

[105] Report of the 2014 Informal Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS) (Report of the 2014 Informal Meeting), UN Doc. CCW/MSP/2014/3, 11 June 2014, para. 1.
[106] *Ibid.*, para. 16. It was stated that "delegations were reminded that the mandate was about emerging technologies".
[107] *Ibid.*, p. 7.
[108] At this point in time, most delegations referred to these weapon systems either as AWS or LAWS, but some delegations, such as Swedish, used these terms interchangeably. For example, "During deployment of any new weapon system, including LAWS, the principles of distinction, proportionality and precautions in attack apply. Despite potential technical advances, of which we have heard examples in the previous sessions of this meeting, it is our assessment that weapon systems will not be capable of performing the subjective evaluations necessary to comply with these rules in the foreseeable future. Targeting decisions and other critical functions, including careful balancing of the principles of IHL, must therefore continue to be made by humans. Today, the technological limitations combined with the rules of IHL, if followed properly, would therefore effectively preclude deployment of an autonomous weapon system." See "Session on Legal Aspects – Swedish Remarks", available at: https://unoda-documents-library.s3.amazonaws.com/ Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_ (2014)/Sweden_Legal%2BAspects_IHL.pdf.
[109] For example, the Holy See used "military robots" or "armed robots" to describe these weapon systems, while Republic of Korea's statement indicated that it was wary of "fully autonomous weapons systems". See "The Use of Lethal Autonomous Weapon Systems, Ethical Questions – Permanent Mission of the Holy See to the United Nations and other International Organisations in Geneva", available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2015)/2015_LAWS_MX_Holy%2BSee.pdf and "Statement by Ambassador Ahn Youngjip", p. 2, available at: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2015)/2015_LAWS_MX_RoK_GS%2BCorr.pdf.

> necessarily inflict death. Others argued that only the lethal use of weapons is relevant in regard to IHL.
>
> 42. While several delegations welcomed the progress made on the conceptual understanding of LAWS, there was a widespread agreement that further work was needed on this issue. In this regard, it was reiterated by many that a GGE, established by the Fifth Review Conference, would be the appropriate body to address a possible working definition."[110]

From this excerpt, it is clear that the informal meetings of experts had never considered lethality as a core or essential component. Rather, it was an issue that warranted further discussion. Therefore, among the recommendations submitted by the Chairperson of the 2016 informal meeting of experts to the CCW Fifth Review Conference was the establishment of an open-ended GGE to identify characteristics and elaborate a working definition for these weapon systems.[111] Subsequently, during the CCW Fifth Review Conference in December 2016, High Contracting Parties to the CCW established the GGE, adopting the recommendation put forth by the 2016 informal meeting of experts.[112]

A review of the GGE's origins and establishment reveals a persistent ambiguity surrounding the essentiality of the term "lethal" in its mandate since 2013. In 2013, when States supported the CCW Chairperson's proposal for an informal meeting of experts, the mandate for this meeting was unclear. This lack of clarity stemmed from the varied descriptions of AWS and the title of the Chairperson's proposal used by different parties, with no audio recording or written documents available to clarify the matter. The meeting report from the 2014 informal meeting of experts further compounded this issue by inconsistent use of terminology.

By 2016, during deliberations among the informal meeting of experts on whether lethality was a required attribute, the group failed to reach a consensus. As a result, they agreed to establish a GGE for further deliberation. Since then, the question of whether the concept of lethality is a necessary requirement has been debated continuously to the present day. Furthermore, reflecting on the GGE's history and debate from 2024 to 2026, while the term "lethal" was included in the name of the group and/or its current mandate, there has been no common understanding of what this term means since the GGE's establishment.[113]

---

[110] Report of the 2016 Informal Meeting, above note 5, pp. 7-8.
[111] *Ibid.*, p. 14.
[112] Final Document of the Fifth Review Conference, above note 3, pp. 13 and 9.
[113] This point was also raised by Brazil in its intervention during the session on 2 March 2026. See "Brazil" (Brazil's Statement at the GGE), available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/83ECCAE6-057E-4EA5-840C-351247F34438/15h03?position=4811.

4. Arguments in Favour for Including Lethality within the Definition and Their Rebuttals

This section provides a brief overview of the arguments for including the term "lethal" in the definition of AWS. States advocating for this inclusion contend that the opportunities afforded by such a classification would allow for a targeted regulatory approach,[114] which aligns with their national security interests, such as force protection and projection. However, the use of weapons in such circumstances does not necessarily require a lethal capability or entail an AWS. Furthermore, while such opportunities purportedly improve compliance with IHL, they largely depend on the specific circumstances in which the weapon systems are used.

Force Protection and Projection

As some States have argued, the primary motivation for including the attribute of lethality is to restrict the scope of the GGE's discussions, ensuring that they encompass only lethal systems.[115] This motivation is ultimately linked to national security interests, including the desires to afford greater protection to soldiers and to extend their reach, thereby enhancing force projection.

With regard to the protection of soldiers, AWS are frequently cited as being suited for the so-called "dirty, dull, and dangerous" tasks.[116] A wide range of military tasks could fall into these categories, from long-duration sorties to explosive ordnance disposal (EOD).[117] However, these tasks do not necessarily require a lethal capability, but rather the use of unmanned platforms. For example, the United States' Unmanned Systems Roadmap: 2007–2032 notes that "ground robots" were used extensively by the coalition forces for EOD efforts in Iraq,[118] and modern EOD operations often renders an explosive device inert without causing it to detonate. Such operations are typically conducted by firing a high-pressure water jet to disrupt

---

[114] For example, China stated that "[a]ddressing the definition of LAWS is the key of negotiating any practical control measures…In our view, we should focus on weapons systems with autonomous killing capabilities, which means autonomous weapons platforms with a lethal mission payload. In terms of lethality, CCW is not intended to prohibit all conventional weapons. Discussions deviated from lethality would not only be contrary to the goal of the CCW, but also make it difficult to have objective and fair discussions on the impacts of Al and other emerging technologies." See "Working Paper of the People's Republic of China on Lethal Autonomous Weapons Systems (July 2022)", above note 44.

[115] *Ibid*; Israel's Statement at the GGE, above note 26.

[116] Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, above note 50, para. 51; Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies*, TMC Asser Press, The Hague, 2024, p. 278; Amitai Etzioni and Oren Etzioni, "Pros and Cons of Autonomous Weapons Systems", *Military Review*, 2017, p. 72, available at: https://www.armyupress.army.mil/Portals/7/military-review/Archives/Eng lish/pros-and-cons-of-autonomous-weapons-systems.pdf.

[117] The United States' Department of Defense, *Unmanned Systems Roadmap: 2007-2032*, Washington DC, 10 December 2007, p. 19, available at: https://www. govinfo.gov/content/pkg/GOVPUB-D-PURL-LPS91893/pdf/GOVPUB-D-PURL-LPS91893.pdf.

[118] *Ibid.*

the device's wires.[119] The use of high-pressure water jet is arguably non-lethal, but it can inflict painful injury which might engage the *maux superflus* rule.[120] On this issue concerning an AWS that seemingly does not possess a lethal capability, it will be further addressed in the section below.

On the other hand, while force projection requires a lethal capability, it does not necessarily entail an AWS. It may be achieved via remotely controlled unmanned platforms.[121] However, it is acknowledged that there are specific circumstances in which AWS equipped with a lethal capability would be particularly useful, including in scenarios where a communication link cannot be maintained. As Thurnher put it, "[lethal autonomous robots] have the unique potential…to lethally strike even when communications links have been severed".[122] These systems may be essential where the operational environment precludes regular communications or where mission success depends on maintaining communication silence, such as in covert operations within contested territory.[123]

Regarding operational environments that preclude regular communications, such as in undersea or underground environments,[124] the use of autonomous weapons could arguably remain compliant with IHL. For instance, in an undersea environment, there are typically few civilian objects, this suggests that such weapons

---

[119] Peter Ray Allison, "What Does a Bomb Disposal Robot Actually Do?", *BBC News*, 16 July 2016, available at: https://www.bbc.com/future/article/20160714-what-does-a-bomb-disposal-robot-actually-do.

[120] Schuller described a hypothetical system named "Surrender Bot", equipped with a high-power laser capable of cutting through an enemy soldier's body armour. He argued that the Surrender Bot is designed to inflict painful injury, but non-lethal, thereby forcing the enemy to surrender. Schuller used this scenario to illustrate a situation in which the *maux superflus* (superfluous injury or unnecessary suffering) rule would apply to AWS. See Alan L. Schuller, "Artificial Intelligence Effecting Human Decisions to Kill: The Challenge of Linking Numerically Quantifiable Goals to IHL Compliance", *Journal of Law and Policy for the Information Society*, Vol. 15, pp. 118-119. For the *maux superflus* (superfluous injury or unnecessary suffering) rule, see Hague Regulations Concerning the Laws and Customs of War on Land (Hague Regulations), 18 October 1907 (entered into force 26 January 1910), Art. 23(e); Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (AP I), 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 35(2); "Customary IHL Database" (CIHL), Rule 70, available at: https://ihl-databases.icrc.org/en/customary-ihl.

[121] J. Kwik, above note 116, p. 278.

[122] Jeffrey S. Thurnher, "No One at the Controls: Legal Implications of Fully Autonomous Targeting", *Joint Force Quarterly,* Issue 67, p. 83.

[123] *Ibid.* One example of such a covert operation is Ukraine's "Operation Spider's Web", conducted in June 2025. While the operation was not fully autonomous, reports suggest that autonomous capability and AI likely played a supporting role in flight navigation, stabilisation, and targeting, enabling the drones to strike their targets within Russia. See Kateryna Bondar, "How Ukraine's Operation 'Spider's Web' Redefines Asymmetric Warfare", *Center for Strategic & International Studies*, 2 June 2025, available at: https://www.csis.org/analysis/how-ukraines-spider-web-operation-redefines-asymmetric-warfare; Katja Bego, "Ukraine's Operation Spider's Web is A Game-Changer for Modern Drone Warfare. NATO Should Pay Attention", *Chatham House*, 6 June 2025, available at: https://www.chathamhouse.org/2025/06/ukraines-operation-spiders-web-game-changer-modern-drone-warfare-nato-should-pay-attention.

[124] J. Kwik, above note 116, p. 283.

could potentially satisfy the principle of distinction.[125] However, suppose that an underwater autonomous weapon was deployed to strike a replenishment tanker of the enemy, it would raise questions regarding the principle of proportionality.[126] This assessment depends on the configuration of the weapon. For instance, whether it is programmed to strike at a specific time and location to minimise environmental impact, or whether it would engage the target immediately upon identification, regardless of the surrounding maritime environment. Similarly, covert operations also raise such IHL issues, especially when the weapon systems were used in urban environments.

Minimise Civilian Harm

Furthermore, while some States have argued that AWS with a lethal capability could offer the potential for greater civilian protection through more precise and accurate strikes,[127] this purported benefit is not primarily achieved through the lethal capability itself, but rather through the autonomous capability of such systems.

Autonomous capability offers a range of benefits, including "faster execution of tasks, higher interoperability, lower error rates, and the increasing use of unmanned platforms, which decreases risk to life and improves operational access".[128] The integration of AI in this regard could arguably help to deliver more precise and accurate strikes, by shortening the time of data processing, which in turn could help an autonomous weapon in executing decisions in a fraction of the time humans could, and ultimately, shortening the time of the OODA Loop (observe, orient, decide, act loop).[129] However, as many experts have argued, there are various difficulties associated with using these weapons to ensure such precision and accuracy.[130]

[125] AP I above note 120, Art. 48; CIHL, above note 120, Rules 1 and 7.
[126] AP I above note 120, Art. 51(5)(b); CIHL, above note 120, Rule 14.
[127] This is also highlighted by Heyns, see Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions—Christof Heyns, above note 50, para. 52.
[128] Andrew P. Williams and Paul D. Scharre, *Autonomous Systems: Issues for Defence Policymakers*, NATO: Headquarters Supreme Allied Commander Transformation, Virginia, 2015, p. 37.
[129] J. Kwik, above note 116, p. 281.
[130] For example, see Noel Sharkey, "Saying 'No!' to Lethal Autonomous Targeting", *Journal of Military Ethics*, Vol. 9, No. 4, 2010; J. Kwik, above note 116, pp. 167-268; Peter Asaro, "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making", *International Review of the Red Cross*, Vol. 94, No. 886, 2012, pp. 696-703, Gary E. Marchant et. al., "International Governance of Autonomous Military Robots", *The Columbia Science and Technology Law Review*, Vol. XII, 2011, pp. 281-289.

5. Counterarguments to the Inclusion of Lethality within the Definition and the Risks of Its Inclusion

This section outlines the counterarguments and risks associated with including the concept of lethality within the definition of AWS. It is found that the arguments offered by some States in opposition to this inclusion are largely grounded in both policy and legal considerations.

Lethality is an Effect of How a Weapon is Used

The first argument against including the concept of lethality within the definition is that lethality is not an inherent design feature of a weapon. Indeed, it is a consequence of how a weapon is used, and therefore, it presents operational challenges in determining whether a weapon system is lethal or not.

Within the GGE, some States have attempted to classify weapons as either lethal or non-lethal.[131] The definition of "non-lethal weapons" is unclear, as the GGE has not attempted to define this term. In scholarly literature, this term is used to encompass weapons that are "sub-lethal", "less lethal", and "less-than-lethal".[132]

The binary classification within the GGE overlooks the fact that a weapon's lethality depends on a wide range of factors, including the manner in which the

---

[131] For example, France and the United States. See Compilation of Replies Received to the Chair's Guiding Questions, UN Doc. CCW/GGE.1/2024/CRP.1, 1 March 2024, pp. 6-7, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_ Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_ Systems_(2024)/CCW_GGE1_2024_CRP1.pdf; "United States", available at: https://conf.unog.ch/digitalrecordings/en/clients/61.0500/meetings/86B33684-84E0-498A-B6B0-327A 27D50405/10h13/en?position=9796. In the United States' intervention during the GGE's session on 5 March 2026, it urged the international community to examine its reservation to the CCW's Amended Protocol II. In this reservation, the United States distinguished between lethal and non-lethal weapon technology. See "The United States of America's Reservation", available at: https://ihl-databases.icrc.org/en/ihl-treaties/ccw-amended-protocol-ii-1996/state-parties/us?activeTab=.

[132] David P. Fidler, "The International Legal Implications of Non-Lethal", *Michigan Journal of International Law*, Vol. 21, 1999, pp. 53-54; David A. Koplow, *Non-Lethal Weapons: The Law and Policy of Revolutionary Technologies for the Military and Law Enforcement*, Cambridge University Press, Cambridge, 2006, pp 9-10. Koplow's book also include weapons that are "disabling" or those that can achieve a "soft kill" or a "mission kill". The term "less lethal" has also been used in literature concerning the use of weapons in armed conflict, including in situations of occupation. See Elisabeth Hoffberger-Pippan, *Less-Lethal Weapons under International Law: A Three-Dimensional Perspective*, Cambridge University Press, Cambridge, 2021. There is also a UN guidance, see Guidance on Less-Lethal Weapons in Law Enforcement, UN Doc. HR/PUB/20/1, 2020. The guidance defines "lethal weapons" as "[w]eapons designed or intended for use on individuals or groups of individuals and which, in the course of expected or reasonably foreseen use, have a lower risk of causing death or serious injury than firearms. Less-lethal ammunition may be fired from conventional firearms. For the purpose of this Guidance, the term includes conventional firearms when they are used to discharge less-lethal ammunition, but not when they are used to discharge conventional bullets or other ammunition that would be likely to result in life-threatening injuries."

weapon is used and the intent of the user. Weapon systems do not have a 100% lethality rate. Weapons which are perceived to be lethal may, in practice, produce non-lethal outcomes. Research has shown that rifles and fragmentation weapons result in only 20-25% of fatalities.[133] Similarly, a study examining the mortality rate associated with weapon use in armed conflict found that the total fatalities never exceeded 26% of all recorded casualties.[134] Conversely, the so-called non-lethal weapons can result in death under certain circumstances.[135] For example, a seemingly innocuous object such as a piece of wood, if used to strike a person repeatedly, can inflict serious injuries, potentially leading to death.

Consequently, lethality is not an inherent characteristic of a weapon, but rather a context-dependent effect determined by the circumstances in which the weapon is used. Any attempt to measure the lethality of a weapon system in absolute terms is inherently ambiguous and unreliable. It also raises questions of whether lethality can be feasibly and practically measured at all.

Non-Lethal AWS May Be Unlawful under IHL

The second counterargument relates to the concern that categorising AWS into lethal and non-lethal may open a regulatory loophole, enabling States to argue that non-lethal AWS fall outside the scope of regulation. States advocating for the distinction between lethal and non-lethal AWS appear to be under the impression that the use of non-lethal AWS is inherently compliant with IHL. For instance, Israel argued that if the GGE does not focus on the use of lethal force, it "risk[s] losing sight of compliance with the key aspects of using weapons under IHL, particularly compliance with targeting rules".[136] However, this line of argument is flawed. Non-lethal AWS are not necessarily compliant with IHL by default.

For instance, blinding laser weapons, typically described as non-lethal, are expressly prohibited under Protocol IV of the CCW because they can cause superfluous injury or unnecessary suffering.[137] Similarly, non-lethal AWS may still be deemed unlawful if they are inherently indiscriminate,[138] or if their use is expected to cause widespread, long-term and severe damage to the natural environment.[139] An

---

[133] Robin M. Coupland, "'Non-lethal' Weapons: Precipitating A New Arms Race", *British Medical Journal*, Vol. 315, 1997, p. 72.
[134] Robin M. Coupland and David R. Meddings, "Mortality Associated with Use of Weapons in Armed Conflicts, Wartime Atrocities, and Civilian Mass Shootings: Literature Review", *British Medical Journal*, Vol. 319, 1999, p. 407.
[135] D. P. Fidler, above note 132, p. 55.
[136] "Statement by Mrs. Michal Maayan, Minister-Counsellor, Deputy Permanent Representative to the Conference on Disarmament", available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Israel's_Statement_-_GGE_LAWS_-_March_4th.pdf.
[137] AP I, above note 120, Art. 35(2); CIHL, above note 120, Rule 70.
[138] AP I, above note 120, Arts. 51(4)(b)-(c); CIHL, above note 120, Rule 71.
[139] AP I, above note 120, Art. 35(3); CIHL, above note 120, Rule 45.

example could include autonomous drone swarms deployed to disperse biological or chemical agents, which, even if not expressly lethal, can cause severe harm to the natural environment and civilian population. As such, categorising AWS according to their perceived lethality risks creating a narrow regulatory focus that is insufficient to address the risks and challenges associated with AWS.

Diverts Attention from the Real Concerns Posed by AWS

The insistence on using the term "lethal" in the definition of AWS diverts attention from the real concerns associated with these systems: their autonomy and the risks and harms they pose. Shifting the focus to lethality risks weakening the regulatory framework for prohibiting or restricting harmful weapons in the future.

Regardless of whether an attack on an enemy soldier or object is sufficiently fatal, the core issue with AWS lies in its inability to carry out the type of abstract reasoning required for IHL compliance. For instance, in assessing the proportionality of an attack, IHL obliges the attacker to determine whether the expected incidental civilian harm would be "excessive" in relation to the anticipated concrete and direct military advantage.[140] This legal assessment is not a simple calculation; it requires complex value judgements, contextual interpretation, and ethical reasoning. From a technical standpoint, such legal assessments are arguably not suitable to be encoded using purely computational or mathematical means.[141] While autonomous weapons may be capable of processing large datasets, they lack capabilities like ethical reasoning and value judgements, which such legal assessments require. [142] Consequently, these systems are fundamentally incapable of making the legal assessments demanded by IHL.

The risk of diverting attention from the actual concerns posed by AWS is that it could ultimately weaken the overall regulatory framework for weapons in general. Traditionally, weapons regulation has focused on their effects, specifically whether they cause unnecessary suffering, superfluous injury, indiscriminate harm, or widespread, long-term and severe damage to the natural environment. Introducing lethality as a criterion for defining a weapon system sets a problematic precedent. In the future, policymakers might be compelled to assess and define a weapon's lethality

[140] AP I, above note 120, Art. 51(5)(b); CIHL, above note 120, Rule 14.
[141] A. L. Schuller, above note 120, pp. 111-121; ICRC and Geneva Academy, *Expert Consultation Report on AI and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts*, Geneva, March 2024, p. 15.
[142] UN Institute for Disarmament Research, *The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems: Background Paper on the Views of States, Scholars and Other Experts*, Geneva, 6 March 2025, p. 24.

before determining whether it should be prohibited and/or regulated.[143] This additional layer of lethality assessment not only introduces technical and operational complexities in determining a weapon system's lethality (since lethality depends on how a weapon is used and the intent of the user as discussed above), but could also delay or hinder timely actions in ensuring the use of such weapons complies with international law.

Conflation of IHL and IHRL Frameworks

Lastly, the concept of lethality is not explicitly recognised in IHL, as argued by some States. Its inclusion in the definition raises concerns that it could blur the distinction between the frameworks of IHL and IHRL, and thereby undermining the legal understanding of warfare, particularly in relation to the legitimacy of the use of force in armed conflict.

While IHL is the primary legal framework under which issues of AWS are being discussed, the framework of IHRL remains relevant because it applies in times of peace and armed conflict.[144] Because IHL provides specific rules on how warfare should be conducted and AWS are mostly used in warfare (at least for now), IHL is the most directly relevant international legal framework for these systems. While distinct, IHL and IHRL both have protective purposes, and their rules may overlap and complement each other.[145] However, these legal frameworks may also have conflicting norms. In such circumstances, the more specific rule prevails over the more general one (to be discussed below).

As mentioned above, the debate surrounding AWS takes place within the IHL framework, specifically the CCW. Within the IHL framework, the concept of lethality is notably absent. This is because the use of force, including lethal force, is presumed to be an inherent aspect of armed conflict, which is the context in which IHL operates. IHL allows lethal force to be directed against lawful targets as a first

[143] The United States has implemented a policy related to the so-called non-lethal weapons. See "DoD Executive Agent for Non-Lethal Weapons (NLW), and NLW Policy", DoD Directive 3000.03E, available at: https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300003p.pdf?ver=2017-09-27-125836-647. Fidler rightly observed that the United States is leading the "non-lethal" weapon movement and warned that the U.S. interpretations of international law might be "self-serving", aimed at promoting the development and deployment of the so-called non-lethal weapons. See D. P. Fidler, above note 132, pp. 95-96.

[144] States have also argued that other branches of international law, such as international criminal law and state responsibility, are also applicable.

[145] Vienna Convention on the Law of Treaties, 1155 UNTS 331, 23 May 1969 (entered into force 27 January 1980), Art. 31(3)(c). For example, while IHL prohibits torture, it does not define it. Therefore, the definition of torture found in the IHRL framework, specifically Article 1 of the Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, can inform its interpretation under IHL.

resort.[146] This is reflected in the core principles of IHL—distinction, proportionality, and precautions—which regulate how attacks must be conducted during armed conflict, rather than whether lethal force should be used in the first place.

In contrast, IHRL imposes much stricter limitations on the use of lethal force. Under IHRL, lethal force may be employed only as a last resort, subject to strict or absolute necessity.[147] This restriction is closely associated with a fundamental concept of IHRL, namely the right to life. But this concept of the right to life is also not explicitly recognised within the IHL framework.

IHRL-based use of lethal force and the right to life are fundamentally at odds with the realities and legal framework governing warfare, where the use of lethal force is not only anticipated but legally permissible when directed against lawful targets under IHL.

According to the doctrine concerning special law repealing general law (known as *lex specialis derogat legi generali*), when two bodies of law offer conflicting norms, the more specific legal regime prevails over the more general one.[148] Since IHL is the body of law specifically crafted to regulate the conduct of hostilities during armed conflict, its provisions take precedence over conflicting rules of IHRL. Therefore, given that IHL does not contain the concepts of lethality and the right to life in the same manner as IHRL, the inclusion of the notion of lethality in the definition of AWS is not only legally unfounded in IHL, but also contradicts the *lex specialis derogat legi generali* doctrine.

Moreover, importing such IHRL standards into the IHL framework risks distorting the legal understanding of warfare, thereby undermining the legitimacy of the use of force in armed conflict. For instance, if IHRL standards were to be applied to a strike against an enemy solider, the strike may be considered unlawful unless it could be shown that no other means, such as capture, were available, and the strike was absolutely necessary. This example underscores the risks of imposing impractical legal obligations on belligerents during armed conflict. It also illustrates the unusual circumstances under which belligerents might be forced to operate, as the use of lethal force would no longer be regarded as a lawful first resort in warfare.

---

[146] "International Humanitarian Law and the Challenges of Contemporary Armed Conflicts in 2015", paras. 168 and 178, available at: https://casebook.icrc.org/case-study/icrc-international-humanitarian-law-and-challenges-contemporary-armed-conflicts-2015.

[147] For example, Article 2(2) of the European Convention on Human Rights provides that lethal force is permissible when it is absolutely necessary to defend any person from unlawful violence, to effect a lawful arrest or to prevent the escape of a person lawfully detained, or in action lawfully taken for the purpose of quelling a riot or insurrection. See Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR), 4 November 1950 (entered into force 3 September 1953), Art. 2(2).

[148] "*Lex specialis derogat legi generali*", available at: https://www.oxfordreference.com/display/10.1093/acref/9780195369380.001.0001/acref-9780195369380-e-1303.

6. Defining "Lethal"

Notwithstanding the counterarguments and risks associated with the inclusion of the concept of lethality within the definition, the GGE has included it and incorporated an interpretation of the term "lethal" into the rolling text.[149] This approach represents a compromise between the two group of States which argue for and against the inclusion of the term "lethal" in AWS' definition,[150] and it can help to overcome the current gridlock. The next question, then, is what the appropriate interpretation of this term is.

As outlined in Section 2, the current state of debate within the GGE is that some States supporting the inclusion of the concept of lethality within the definition were seeking further concessions from those that opposed, and to limit the interpretation of the term "lethal" to encompass only human fatalities.

In finding an appropriate interpretation, one should examine the term's ordinary meaning, considered within its specific context and in accordance with the relevant instrument's object and purpose.[151] While acknowledging that the rolling text is not a legally binding treaty, and that the GGE does not have a mandate to negotiate a legally binding treaty for these weapon systems, the interpretive methodology contained herein is intended to move the discussion away from States' preferences, and to better align with the objectives of the rolling text.

If one adopts the *Oxford English Dictionary* definition, the term "lethal" can be understood as being "sufficient to cause death", or "very harmful, destructive".[152] Consequently, a definition of lethality that solely focuses on human fatalities does not sit neatly with this ordinary meaning. Furthermore, such a definition would also fail to account for the effects of these weapon systems on objects. They are not exclusively used against humans. If the definition of lethality is confined solely to human fatalities, it would not only neglect the impacts AWS can have on objects, but also create an internal inconsistency within the rolling text. For example, within the rolling text, there is an element limiting or restricting the use of these weapon systems to objects that qualify as military objectives by nature (as seen in the third version of the rolling text)[153] or to military objects (as seen in the fourth version of the rolling text).[154]

---

[149] As seen in the fourth iteration of the rolling text. However, this may change in future debates within the GGE.

[150] The GGE Chairperson's summary of the March 2026 meeting noted that "[t]he formulation in sub-paragraph 1B appeared to represent a compromise that accommodates these differing perspectives." See Chair's Summary – First 2026 session of the GGE on LAWS, UN Doc. CCW/GGE.1/2026/WP.2, 1 April 2026, para. 14.

[151] Vienna Convention on the Law of Treaties, above note 145, Art. 31(1). This provision provides that "[a] treaty shall be interpreted in good faith in accordance with the ordinary meaning to be given to the terms of the treaty in their context and in the light of its object and purpose."

[152] "Lethal", available at: https://www.oed.com/dictionary/lethal_adj?tab=meaning_ and_use.

[153] Rolling Text - 12 May 2025, above note 10, p. 3.

[154] Rolling Text - 18 December 2025, above note 11, p. 3.

Moreover, even if the proposals were amended to include damage and destruction of objects—while still excluding human injury—they would remain at odds with one of the core principles of IHL: the proportionality principle. According to this principle, an attack is considered disproportionate if it is "expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."[155] Notably, this principle does not limit its scope to human fatalities; it includes human injuries as well. Consider a scenario in which an autonomous system targeting a human has failed to produce fatal effects and has caused injuries to a group of civilians in the vicinity of the target. Such an incident would raise questions, such as whether those injuries suffered by the civilians would fall under the proportionality principle. These are precisely the types of legal ambiguities that would arise if such narrowly framed definitions of lethality were to be adopted. Additionally, the proposals seeking to interpret "lethal" to encompass only human fatalities overlook the fact that weapons cannot guarantee a 100% lethality rate. The lethal effect of a weapon depends on the circumstances in which it is used.

In contrast, the proposals for defining lethality that include not only human death and injury, but also the destruction, damage or neutralisation of objects more accurately reflect the diverse effects of these weapon systems. By taking these diverse effects into account, this broader definition of lethality helps to close two potential regulatory loopholes. First, it ensures that the effects of AWS on objects are not overlooked. Second, it acknowledges that the so-called non-lethal AWS may still fall within the regulatory bounds.

7. Conclusion

Mapping and analysing the issue of lethality within the GGE under its current mandate—from the initial debate concerning whether to include the concept, to the proposed definitions and the responses from States—reveals that some States that insisted on including the concept of lethality within the definition of AWS have adopted an inconsistent position. On the one hand, they maintained that the term "lethal" is central to the GGE's mandate and necessary to limit its scope of discussions. However, their general support for the proposed definition and its clarifying sentence (as stated in the third version of the rolling text)[156] or the clarifying sentence (as seen in the fourth version of the rolling text),[157] which would expand that scope, deviates from their original rationale. While this shift in position could be perceived as a constructive way forward for the two groups of States, it casts doubt on the underlying reasons for their initial insistence.

---

[155] AP I, above note 120, Art. 51(5)(b); CIHL, above note 120, Rule 14.
[156] As discussed in Section 2. Also see Chair's Summary – Second 2025 Session of the GGE on LAWS, above note 46, para. 12.
[157] Japan's Statement at March 2026 Meeting, above note 46; See Chair's Summary – First 2026 session of the GGE on LAWS, above note 150, para. 14.

The key factors driving some of these States in supporting the inclusion of the concept of lethality within the definition are mainly related to national security interests, including force protection and projection. An examination of these interests, however, reveals that the use of autonomous weapons for both force protection and projection does not necessarily require a lethal capability and/or entail a specific AWS. These objectives could be achieved by other means, such as the use of remotely controlled unmanned systems. More importantly, even if an AWS does not possess a lethal capability, its use may still raise IHL concerns.

This is because weapons do not have a 100% lethality rate. A so-called lethal AWS may produce non-lethal effects, while a non-lethal AWS may cause lethal effects. Therefore, an artificial separation of lethal and non-lethal AWS (or any other classifications such as "less lethal", "less-than-lethal" or "sub-lethal" as mentioned in Section 5) raises questions regarding the practicality and feasibility of determining a weapon's lethality. Additionally, since IHL allows the use of force as a first resort (unlike IHRL), the necessity for such an artificial classification is unclear, unless it serves as a means to limit the scope of the regulatory framework, and carve out a space for the continued development and use of the so-called non-lethal AWS in warfare.[158]

Furthermore, while some States contend that the attribute of lethality is necessary because it is explicit in the GGE's mandate, an examination of the group's origins shows that the informal meetings of experts had never considered lethality as a core or essential component. Rather, it was an issue that warranted further discussion. This subsequently led to the establishment of the GGE in 2016, which was tasked with identifying the characteristics and elaborating on a working definition of these weapon systems. Thus, whilst the name and mandate of the group include the term "lethal", there has been no common understanding of this term since its inception.[159] Consequently, the issue of lethality has remained a subject of debate long after the GGE's establishment, continuing to the present day.

---

[158] Similar arguments have been in relation to the so-called non-lethal weapons, but not in the context of non-lethal AWS. For example, see D. P. Fidler, above note 132, p. 96. Another reason relates to defence industry interests. Some States that are insisting on the inclusion of the concept of lethality within the definition are the world's leading arms exporters. For example, the United States' share of global arms exports currently stands at 42%, Russia at 6.8%, China at 5.6%, Israel at 4.4%, and Japan at 3.0%. These States are the top 25 largest arms exporters. See SIPRI, *Fact Sheet: Trends in International Arms Transfers 2025*, Solna, March 2026, p. 2. For a brief discussion on this issue on defence industry interests, see Mei Ching Liu and Manoj Harjani, "International Regulation of Lethal Autonomous Weapon Systems: Divergence and the Way Forward", *IDSS Paper*, 30 March 2026, available at: https://rsis.edu.sg/rsis-publication/idss/ip26056-international-regulation-of-lethal-autonomous-weapon-systems-divergence-and-the-way-forward/. On the other hand, Fidler identified five key factors in driving the interest in the concept of "non-lethal" weapons: (i) technological developments; (ii) changes in the roles of military forces; (iii) pressure to retain competitive advantages against adversaries; (iv) moral and legal pressure exerted by international law on armed conflict; and (v) marketing by the military. See D. P. Fidler, above note 132, pp. 57-60.

[159] Brazil's Statement at the GGE, above note 113.

The issues surrounding the concept of lethality within the definition are not merely semantic. They carry significant implications for international peace and security. The definition of AWS would directly shape the scope of a future regulatory framework. Including the term "lethal" in the definition would narrow the regulatory framework, meaning that only AWS capable of producing so-called lethal effects would fall within the framework. This could significantly limit the types of autonomous systems subject to regulatory oversight. Conversely, excluding the term "lethal" from the definition would prevent artificial restrictions on the framework. It would allow for broader regulatory coverage and limit the development and use of AWS that would, in some or all circumstances, be prohibited by international law.

However, as discussed in this article, there is a provisional agreement from States that originally opposed the inclusion of the concept of lethality. These States appear willing to compromise. However, those seeking its inclusion were demanding further concessions, specifically that the term "lethal" be interpreted solely as encompassing human fatalities. To overcome the divergences related to the interpretation, it is recommended that a broader definition of lethality be adopted, which encompasses not only human fatalities but also human injuries, as well as the damage, destruction and neutralisation of objects.

Nevertheless, it is recommended that the concept of lethality should ultimately not be embedded in any future final legal instrument regulating AWS (provided that the GGE receives a mandate to commence such negotiations or if such negotiations are initiated through other platforms, such as the UN General Assembly). States' concern regarding the term's lack of a basis in IHL has far greater implications, extending beyond the AWS regulatory framework.

Incorporating the concept of lethality into a weapon regulatory framework not only signifies a departure from how weapons are traditionally governed, but also contradicts the *lex specialis derogat legi generali* doctrine. The implication of contradicting this legal doctrine is that it could undermine the established coherence and hierarchy of international legal principles, making it more challenging for States to determine which legal norms apply in various situations. It could also blur the distinction between the frameworks of IHL and IHRL, thereby altering our understanding of warfare. Lastly, the concerns with incorporating the concept of lethality into a legally binding instrument lie not only in the technical and operational difficulties of assessing a weapon's lethality, but also in the risk that this additional layer of lethality assessment could be used to undermine the broader weapons regulatory framework.

In light of the above, the international community should resist efforts that could blur the distinction between IHL and IHRL, as doing so could set a dangerous precedent for how weapons are regulated and how the use of force in armed conflict is interpreted under international law.

PART 4
LEGAL RESPONSIBILITY FOR MILITARY AI

Transcending the Dilemma of Applying IHL to LAWS in Agency Spillover Engagements: A Legal Analytical Perspective on Human-Machine Hybrid Agency

Fan Yang[*]

ABSTRACT

The militarization of Lethal Autonomous Weapons Systems inevitably triggers "agency spillover engagements", situations where algorithmic agency deviates from human presets. This has led to systemic applicability ruptures across International Humanitarian Law, state responsibility, and International Criminal Law. This paper argues that the root cause of the failure lies in the strict adherence to the human moral agent presumption. To address this, this paper proposes an ontological perspective of human-machine hybrid agency and employs the principal-agent relationship as an analytical lens. It further derives three domain-specific tools – the Joint Cognitive System, Risk Allocation, and Reckless Delegation – to transcend these dilemmas within lex lata *through evolutionary and teleological interpretation.*

Keywords: Lethal Autonomous Weapons Systems; Agency Spillover Engagement; Hybrid Agency; Principal-Agent; International Humanitarian Law; Legal Interpretation

I. Introduction

Compared to all previous weapon systems, Lethal Autonomous Weapons Systems (LAWS)[1] exhibit one fundamental difference: due to the intervention of Artificial Intelligence (AI), LAWS intrinsically possess a certain degree of autonomous agency.[2] Similar to civil or commercial AI systems that have rapidly evolved over

---

[*] Associate Professor of International Law, Deputy Director of Cyberspace International Law Center, Xiamen University School of Law, China.

[1] The scope of military AI is analytically broader than LAWS, including using AI capability in military cyber operations, as well as AI-enabled decision support system in military operations. See Fan Yang, *Applying International Law to AI in the Military Domain: An Integrated Approach*, GC REAIM Expert Policy Note Series, April 2025, available at: https://hcss.nl/wp-content/uploads/2025/04/Yang.pdf. (all internet references were accessed on May 6, 2026). This paper's focus is on LAWS.

This paper retains the term "LAWS" as it remains the established designation in the CCW GGE process. The author acknowledges ongoing debate about whether the qualifier "lethal" is clearly defined or analytically necessary; the framework proposed here applies to autonomous weapon systems with lethal consequences regardless of the terminological convention adopted.

[2] The distinction between "autonomous" and "automatic" (or "automated") is foundational to defining LAWS. A merely automatic system follows predetermined, deterministic rules; an autonomous system can modify its behaviour in response to environmental conditions. See Paul Scharre, "*Army of None: Autonomous Weapons and the Future of War*", W.W. Norton, New York, 2018, pp. 27-45; see also United States Department of Defense, Directive 3000.09, "Autonomy in Weapon Systems", 25 January 2023, defining autonomous weapon systems as those that "can select and engage targets without further

the past two years (e.g., agentic AI, autonomous driving, embodied intelligent robots),[3] these AI weapon systems possess an "entity capacity" to autonomously plan, interact, and alter the state of reality in digital or physical environments with minimal human intervention.[4] The law needs to properly incorporate this agency into its analytical framework, just as it has historically succeeded in addressing the endogenous moral agency of humans, the biological agency of animals,[5] and the fictive group agency[6] of organizations such as corporations and partnerships. Ignoring it or failing to assimilate it will inevitably lead to legal dilemmas. This is because AI agency can disrupt or even sever the legal analysis of human responsibility that would apply in equivalent or similar situations without AI. Human subjects will frequently argue as a defence that, under the current legal framework, liability cannot be attributed to them because the harmful consequences were caused by AI agency.

An ideal research conclusion concerning LAWS and International Humanitarian Law (IHL) should be able to effectively address this agency,[7] analysing the determination of illegality and the pursuit of legal responsibility resulting from the use of LAWS during engagements – namely, whether there is a violation of IHL rules on the conduct of hostilities, and whether it triggers state responsibility and/or individual international criminal responsibility. Regarding the issue of legal responsibility, the scope of legal analysis transcends the specific branch of IHL, entering into the realms of the Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA) or International Criminal Law (ICL). However, these inquiries remain inextricably linked to, or premised upon, the IHL analysis.

---

intervention by a human operator." For a comparative analysis of definitions, see Filippo Santoni de Sio and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account", *Frontiers in Robotics and AI*, Vol. 5, No. 15, 2018.

[3] Many general discussions on the legal liability of AI also choose to approach the issue from the perspective of agency. See, e.g., Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay and Christoph Winter, "Law-Following AI: Designing AI Agents to Obey Human Laws", *Fordham Law Review*, Vol. 94, 2025, pp. 57-129.

[4] Luciano Floridi defines agency through three dimensions: the capacity to interact with and affect the environment (interactivity), the capacity to change state independently of external control (autonomy), and the capacity to change its own rules of behaviour based on input (adaptability). See Luciano Floridi, "AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis", *Philosophy & Technology*, Vol. 38, 2025, p. 30.

[5] See L. Floridi, above note 4, p.30.

[6] See Christian List, "Group Agency and Artificial Intelligence", *Philosophy & Technology*, Vol. 34, 2021, pp. 1213-1242.

[7] Military AI decision-support systems also present agency issues, but these primarily concern how human agency is reshaped or even eroded following the introduction of such new technologies. For research specifically discussing how this impact reflects on the legal analysis of IHL applicability. See Taylor Kate Woodcock, "Human/Machine(-Learning) Interactions, Human Agency and the International Humanitarian Law Proportionality Standard", *Global Society*, Vol. 38, No. 1, 2024, pp. 100-121.

Logically speaking, the legal crises that genuinely challenge the application of IHL in a dogmatic sense – which may be a factual reality or a defence invoked by potential responsible parties for exculpation – can be narrowed down to a specific category of scenarios. Due to the underlying agency of the weapon's AI system, compounded by complex interferences in real battlefield environments, LAWS may act beyond the preset instructions and rules of engagement (ROE) of human commanders or operators, resulting in consequences that allegedly violate IHL. This paper terms such scenarios Agency Spillover Engagements (ASE).[8] Conversely, when LAWS appear to follow the preset instructions and ROE of human commanders or operators, their inherent agency overlaps with, or can be deemed subsumed by, the agency of the relevant human subjects. Consequently, legal analysis needs only to address the latter, rendering the entire legal reasoning identical to traditional analytical pathways.

Examining the agency of LAWS and human subjects along a continuous spectrum helps explain numerous phenomena. For instance, the emphasis on ensuring "meaningful human control" in the context of LAWS,[9] or the demand for enhanced explainability during the research and development and design phases of weapon AI systems,[10] and other propositions like these all point toward a unified endeavour: the desire to institutionally limit and narrow the exercise of LAWS' agency, while expanding the dominance and suppression by human agency. Similarly, the consensus advocating for the prohibition of fully autonomous weapons[11] represents the international community's shared resistance against the extreme scenario where LAWS' agency completely marginalizes human agency. Despite these efforts, it must be acknowledged that as long as LAWS are introduced to the battlefield, the possibility of an agency spillover engagement inevitably exists. This anxiety does not stem from anthropomorphic imaginations found in sci-fi narratives of machines awakening and breaking free from human control; rather, it

[8] Some scholars refer to this as "attacks beyond preset parameters (in Chinese 超设定攻击)."See LENG Xinyu, "Principle of Meaningful Human Control under Topic of Lethal Autonomous Weapon System", *Journal of International Law (China)* Vol. 2, 2022, p. 17.

[9] For example, the principle of meaningful human control is expressed as making the utmost effort to implement human control over weapon systems during policy formulation, development, deployment, testing, use, and post-use assessment. The 2018 GGE report labelled these processes with numbers 0 to 5 and depicted them in a chart to facilitate understanding and policy adjustment by states. See Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2018/3, Annex III, para. 10.

[10] See, e.g., International Committee of the Red Cross. (2021). *ICRC position and background paper on autonomous weapons systems*. International Committee of the Red Cross: Technical report, p. 7.

[11] This consensus is reflected in the "rolling text" of the Group of Governmental Experts (GGE) on LAWS under the Convention on Certain Conventional Weapons (CCW), as well as in the policy positions of major military AI powers and representative military AI enterprises. See, e.g., the Report of the 2023/2024 Session of the GGE on LAWS, and the *Blueprint for Action* endorsed at the 2024 REAIM (Responsible AI in the Military Domain) Summit.

is founded upon a stark technological reality.[12] When a system based on complex algorithms is granted a certain degree of autonomous decision-making authority in an open, chaotic battlefield environment, it will – when faced with unpredictable environmental stimuli – inevitably have a probability of generating emergent actions based on its underlying computational logic that a human commander could not predict *a priori*.

Consider the following hypothetical scenario: an autonomous loitering munition deployed for Suppression of Enemy Air Defences (SEAD) encounters an unknown electronic deception tactic. Instead of conventionally choosing to self-destruct or return to base, its algorithm autonomously learns and determines that a nearby civilian communication tower is a disguised interference source, subsequently initiating an attack to destroy it. The loitering munition does not experience a mechanical malfunction in the traditional sense; rather, while exercising its delegated authority, it executes an agency spillover engagement unforeseen by humans.

At this juncture, the core substantive obligations of IHL face the disintegration of their applicative foundation due to the absence of human empathy and subjective moral balancing. Simultaneously, because the attack deviates from specific State authorization and the direct intent (*mens rea*) of the commander, the attribution chain of State responsibility and the individual accountability mechanisms of International Criminal Law are dually obstructed.[13]

Since agency and specifically, its spillover, are the culprit creating this legal analytical dilemma, agency might also hold the key to resolving it. This paper attempts to use the agency of LAWS as a continuous thread of inquiry, drawing on both legal philosophy and philosophy of technology. The logical progression of this paper is as follows: First, it systematically examines the specific legal obstacles facing

---

[12] This is a technological inevitability. On the technical mechanisms through which machine learning systems produce unintended and harmful behaviour, including distributional shift (encountering environmental data outside the training distribution), reward hacking, and negative side effects, see Dario Amodei *et al.*, "Concrete Problems in AI Safety", arXiv preprint 1606.06565, 2016. For analysis of how these risks manifest specifically in lethal autonomous weapons systems, see Ingvild Bode and Tom Watts, "Technical Risks of (Lethal) Autonomous Weapons Systems", arXiv preprint 2502.10174, 2025, arguing that even rigorously tested systems may behave unpredictably and harmfully in real-world conditions due to the black-box nature of AI decision-making, susceptibility to reward hacking, goal misgeneralisation, and emergent behaviours that escape human control. See also Dustin Trusilo and David Danks, "Autonomous AI Systems in Conflict: Emergent Behavior and Its Impact on Predictability and Reliability", *Journal of International Humanitarian Legal Studies*, Vol. 14, No. 2, 2023, discussing how autonomous algorithmic systems operating in open environments produce nondeterministic, innovative and nontransparent behaviour at the level of data processing.

[13] See Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technology*, Vol. 6, 2004, pp.175-183. The academic community has further subdivided this gap into the culpability gap, moral accountability gap, and active responsibility gap. See Filippo Santoni de Sio and Giulio Mecacci, "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them", *Philosophy & Technology*, Vol. 34, 2021, pp.1057-1084.

IHL substantive rules and accountability mechanisms in agency spillover engagements, revealing the limitations of current interpretative pathways. Second, it traces and reconstructs the concept of agency in legal history and jurisprudence, arguing that during engagements utilizing LAWS, the functional agency of the AI-driven weapon system combines with the agency of human subjects to constitute a hybrid agency, which is precisely the core object that legal systems such as IHL must properly incorporate into their analysis. Finally, by developing conceptual tools related to agency, the paper explores how to transcend existing legal analytical barriers.

II. The Triple Legal Dilemma in Agency Spillover Engagements

In the ASE scenario, strictly adhering to the analytical pathways of legal dogmatics inevitably drives the substantive rules of IHL on the conduct of hostilities, the attribution logic of ARSIWA, and the individual accountability mechanisms under ICL into an impasse. Such agency spillover can be subsumed neither by the traditional legal concept of weapon malfunction nor by the doctrine of mistake of fact.

A. The Applicability Dilemma of Core Substantive IHL Rules

Under the dogmatic framework, the bearers of substantive IHL obligations are consistently human commanders or operators, a consensus that the international society has upheld till now.[14] The true legal crisis posed by LAWS in ASE lies in the fact that, the factual agency they exhibit substantively severs the traditional evaluative chain for determining whether a human subject has violated IHL obligations. Many IHL rules will have to answer this challenge, but here we focus on the three IHL principles on the conduct of hostilities.

The principle of distinction requires conflicting parties to distinguish between civilian and military objectives at all times.[15] In the traditional use of weapons, there exists a linear correspondence between a commander's intent to attack and the weapon's physical point of impact. However, deploying LAWS means outsourcing the cognitive process of target feature recognition and classification at the tactical edge to algorithms. When LAWS, operating in a complex electromagnetic environment, autonomously generate a decision to classify a civilian facility as a military objective and subsequently destroy it, the human's lawful initial intent and the unlawful final attack result are decoupled by the machine's agency.

[14] See Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1./2019/3, Annex 4, p.14.
[15] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 48. (AP 1). "… the Parties to the conflict shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives."

Consequently, current IHL is unable to subsume this scenario within its traditional framework to hold the human commander in violation of the principle of distinction.

The principle of proportionality similarly faces the disintegration of its spatiotemporal foundation. This principle heavily relies on the decision-maker's subjective balancing of the "anticipated military advantage" against the "incidental damage" at a specific moment.[16] A commander's general estimation upon initiating deployment in the rear is vastly disconnected from the specific attack trajectory instantaneously generated by LAWS deep within the battlefield. Confronted with a spillover attack scenario temporarily created by machine agency, the commander arguably never conducted any "balancing" in the IHL sense, causing the evaluative chain to rupture at this exact juncture.

The principle of precautions requires the attacker to "cancel or suspend" an attack if it becomes apparent that the objective is not a military one.[17] This obligation is premised on the human possessing the physical and cognitive feasibility to comply. However, this faces an impenetrable physical barrier at the tactical end, as machine algorithms complete the OODA (Observe-Orient-Decide-Act) loop in milliseconds. When an ASE occurs, the commander is cognitively unable to penetrate the algorithmic black box and is physically excluded from the decision-making loop. Continuing to compel the commander to "suspend the attack in a timely manner" would contradict the fundamental legal maxim of *ultra posse nemo obligatur*, i.e. no one is obligated beyond what they are able to do.

B. The Applicability Dilemma of Attribution Rules in State Responsibility

The violation of IHL substantive obligations may lead to State responsibility, provided that such a violation can be attributed to a specific State under international law.[18] Arguably, the attributability of an IHL violation caused by the use of LAWS is self-evident. According to Article 4 of ARSIWA, an attack conducted by the armed forces of a State using any weapon system is naturally and directly attributable to the State, regardless of whether it uses a rifle, a missile, or an AI weapon. Therefore, there is no need to invoke other rules of attribution; the legal attribution judgement in the sense of State responsibility is clear and unquestionable.

---

[16] AP I, Art. 51(5)(b). Indiscriminate attacks are prohibited, including "an attack which may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."
[17] API, Art. 57(2)(b). "An attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated".
[18] ARSIWA, Art. 2.

In the ASE scenario, a State might construct the following defence.[19] The attack by LAWS did indeed occur; however, the act is a manifestation of the weapon system's own agency, which constitutes an "independent intervening factor" (*novus actus interveniens*). Its connection to the operator is limited to the fact that the operator activated the weapon system, a fact that is a condition for the IHL violation rather than the proximate cause.[20] Since it cannot be regarded as the result of the operator's conduct, how can it then be attributed to the State? There are three alternative analytical approaches, each of which encounters interpretive difficulties that illuminate the broader challenge LAWS pose to the existing framework.

The first path attempts to attribute the autonomous conduct of LAWS directly to the State. Under Article 5 of ARSIWA, the conduct of a "person or entity" exercising elements of governmental authority shall be considered an act of the State. According to the International Law Commission's commentary, authorized entities can encompass a wide range of forms, such as public corporations, semi-public entities, and private companies, under the crucial condition that they are empowered by the internal law of the State.[21] Although LAWS objectively exhibit functional agency and can be considered to have been empowered to a certain extent, thus enabling them to autonomously search, identify, and engage targets; the critical issue is that, at least at the current stage, LAWS are not recognised as possessing legal personhood, and they cannot be directly interpreted as the "entity" referred to in Article 5.[22]

Then, could one invoke Article 8 of ARSIWA to treat the conduct of LAWS as conduct "directed or controlled" by a State? The application of Article 8 strictly requires the State to exercise "control" over the specific conduct. But in ASE, LAWS autonomously generate an attack trajectory beyond preset parameters based on underlying algorithms in a complex environment. At this point, the factual agency of the machine has transcended the State's initial boundaries of authorization. The State could argue in defence that the transgressive attack is neither the direct will of the State apparatus nor within the State's instantaneous control, but purely the result of the independently emergent agency of the algorithmic agent. Not to mention, the

---

[19] Some scholarly view holds that such a defence would not succeed for State-deployed autonomous capabilities. See Jonathan Kwik, "Autonomous Cyber Capabilities under International Law", CCDCOE (2021), p. 128; see also Heather A. Harrison Dinniss, "AI and Military Cyber Operations", in *Research Handbook on Warfare and Artificial Intelligence*, Edward Elgar Publishing, 2024, p. 274. This paper acknowledges and agrees with that conclusion. The defence is nevertheless constructed here in its strongest form, because doing so reveals precisely *where* the existing doctrinal framework struggles and *why* the hybrid agency perspective is needed to provide the theoretical grounding for the rejection.

[20] James Crawford, *State Responsibility*, Cambridge University Press, Cambridge, 2013, pp. 492-494.

[21] ILC Commentaries to the draft articles on Responsibility of States for internationally wrongful acts, extract from the Report of the International Law Commission on the work of its Fifty-third session, Official Records of the General Assembly, Fifty-sixth session, Supplement No. 10 (A/56/10), chp.IV.E.2, p. 92.

[22] For discussion that opposes anthropomorphizing LAWS, see, e.g., Bill Boothby, "AI Warfare and the Law", *International Law Studies*, Vol. 104, 2025, pp. 9-12.

interpretative standards of "effective control" [23] or "overall control" [24] have historically been used to address the relationship between States and armed opposition groups that are human organizations. Forcibly creating a legal fiction of AI as a highly organized non-State entity to mechanically apply Article 8 is a category mistake.

Another path attempts to build a bridge through the human operator. Since the operator belongs to the State's armed forces, falling under the definition of a State organ in Article 4 of ARSIWA, the State should bear responsibility as long as it is determined that it is the "conduct of" the operator that LAWS carries out an unlawful attack. However, in ASE, this transmission chain is similarly severed. When LAWS autonomously lock onto a target with millisecond decision-making speed and opaque black-box algorithms, the high-frequency agency of the machine substantively deprives the operator of cognitive and interventional capabilities at the tactical end. The operator can neither foresee the specific unlawful attack nor easily implement an effective suspension physically. As such, the unlawful engagement behaviour cannot be simply and directly evaluated in law as the conduct of the operator. The difficulty lies in determining the standard by which an autonomous weapon's attack beyond preset parameters is still considered that of the operator. Furthermore, whether the operator is responsible as a principal offender for the autonomous weapon's unlawful attack depends on an analysis of their subjective state. In other words, this analytical path entangles the question of State attribution with the determination of individual criminal responsibility. This pushes the legal analysis into the next thorny dilemma.

C. The Applicability Dilemma of Individual Accountability in ICL

Assuming that the autonomous weapon system itself does not involve a weapon whose use is prohibited or restricted by international law,[25] the problems arising in ASE manifest in two aspects when ICL attempts to pursue the individual criminal responsibility of relevant human subjects. The first is whether a causal link can be established between the result of the weapon system's attack violating IHL and the operator's conduct. The second is how to evaluate the psychological state of the operator. This corresponds to ICL's fundamental method of combining objective and subjective elements to attribute guilt.

---

[23] International Court of Justice, *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Judgment (Merits), ICJ Reports 1986, para. 115; International Court of Justice, *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro)*, Judgment, ICJ Reports 2007, para. 400.

[24] International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Duško Tadić*, Case No. IT-94-1-A, Judgment (Appeals Chamber), 15 July 1999, paras. 116–141.

[25] When an autonomous weapon system involves a weapon whose use is prohibited by international law, the mere use of the autonomous weapon system already constitutes a war crime. These war crimes are conduct crimes, where a conviction can be secured without the need to consider the criminal outcome or the causal relationship that caused the criminal outcome.

Particularly concerning the subjective elements, LAWS create a degree of a responsibility vacuum. In ASE, although the commander has pressed the button to activate the system, which constitutes an objective act of deployment, they may not possess "knowledge" in the criminal law sense regarding the subsequent transgressive attack autonomously generated by the machine agent, let alone the "intent" to pursue such an unlawful consequence. Here, a rupture exists between the human's criminal intent (*mens rea*) and the machine's actual harmful active conduct (*actus reus*). It is difficult to constitute direct intent, amounting at most to indirect intent.

Per ICL jurisprudence, the subjective elements of war crimes generally encompass direct intent, with indirect intent existing only as an exception. International criminal justice has no precedent for holding a principal perpetrator criminally liable based on negligence, and only in limited circumstances does it recognise a situation where the defendant in a war crime case possessed the mental state of indirect intent.[26] Article 30 of the Rome Statute sets extremely stringent requirements for the subjective elements of war crimes, requiring the perpetrator to act with "intent and knowledge". It requires that the person "means to engage in the conduct" and "is aware that a consequence will occur in the ordinary course of events." The jurisprudence of the International Criminal Court (ICC) has adhered to the explicit limitations contained in Article 30 of the Rome Statute, thereby excluding indirect intent. The mainstream view continues to follow Article 30, limiting criminal intent to first-degree direct intent (*dolus directus* of the first degree) and second-degree direct intent (*dolus directus* of the second degree). Therefore, in the absence of clearer evidence on State practice proving that the subjective elements of war crimes universally include indirect intent, it is unfeasible, under the direct intent standard as currently applied by the ICC, to pursue the operator's criminal responsibility as a principal perpetrator based on an unlawful attack by an autonomous weapon that exceeds the operator's preset parameters.

That said, the direct intent standard is not the only possibility. Some national jurisdictions and international tribunals do recognise recklessness or *dolus eventualis* as a sufficient *mens rea* for war crimes.[27] This raises the question of whether a recklessness pathway might resolve the ICL dilemma. However, even if recklessness

---

[26] See, e.g., Mucić et al. ("Čelebići") Trial Judgment, 16 November 1998, para. 437; Blaškić Trial Judgment, para. 153; Tadić, Appeal Judgment, 15 July 1999, para. 232; Stakić Trial Judgment, 13 July 2003, para. 587; Brđanin Trial Judgement, *ibid*, para. 589; Naletelić Trial Judgement, para. 577; Kordić and Čerkez Trial Judgement, para. 341; Strugar Appeal Judgement, para. 277; Prosecutor v. Strugar ("Dubrovnik"), Appeal Judgment, IT-01-42-A, 17 July 2008, para. 270; Prosecutor v. Galić, Judgment, IT-98-29-T, 5 December 2003, para. 55.

[27] The Rome Statute is not the exclusive source for determining the mental state required for war crimes liability; customary international law may set different standards, which could arguably be accommodated through Article 30's "unless otherwise provided" clause in International Criminal Court Statute. See Gerhard Werle and Florian Jessberger, "'Unless Otherwise Provided': Article 30 of the ICC Statute and the Mental Element of Crimes under International Criminal Law", *Journal of International Criminal Justice*, Vol. 3, 2005, pp. 35-55.

is accepted as the applicable mental state, two structural problems remain. First, recklessness must attach to a specific criminal act, and without reconceiving the deployment authorization, rather than the machine's autonomous strike, as the relevant act, there is no conduct to which the commander's reckless state of mind can be connected. Second, establishing a causal link between the commander's reckless decision and the specific downstream harm requires bridging the algorithmic black box, which conventional proximate causation analysis cannot accomplish. Overcoming these structural problems requires fresh perspective and analytical framework.

Beyond this, theoretically, one might attempt to stylize LAWS as a criminal subject and seek to pursue the responsibility of the humans behind them through "Joint Criminal Enterprise" (JCE) or "Command Responsibility." [28] Even disregarding the previously mentioned consensus against anthropomorphizing LAWS into legal subjects,[29] this argumentative path faces fundamental obstacles.

Faced with algorithmic entities possessing a high degree of autonomous agency, epistemological barriers render the commander unable to penetrate the black box of LAWS to "know" exactly when and where the system will spill over.[30] JCE requires participants to share a "common purpose" and a "meeting of minds".[31] The *ad hoc* tribunals further clarified that a defendant need not have foreseen an excess act by a group member as inevitable; it suffices that the act was a "possible consequence" of executing that common purpose. In this context, the conundrum lies in how to explain that both the autonomous weapon and the operator or their commander recognise the existence of a shared "common purpose," as well as the autonomous weapon's volition to advance such a common purpose.

The subjective elements for establishing command responsibility are that the defendant, acting as a superior, knew or had reason to know that their subordinates were committing or about to commit crimes. [32] Both customary law and ICC jurisprudence confirm that the legal standard for determining a superior-subordinate relationship is whether the commander personally exercises "effective control" over

[28] For a recent comprehensive analysis of JCE and command responsibility challenges in the context of autonomous weapon systems, see Antonio Coco, "Modes of Liability for AI-Enabled Crimes in International Criminal Law", *International Law Studies*, Vol. 107, 2026, pp.228-255.
[29] Regarding this methodological approach, a possible counter-argument exists: although the law does not recognise autonomous weapons as having subject status, for the necessity of applying existing international criminal law and State responsibility rules, autonomous weapons are treated as fictive subjects. The purpose is not to pursue the responsibility of the fictive subject, but to focus on pursuing the responsibility of individuals and States as subjects of international criminal law.
[30] Prosecutor v. Tadić, Appeals Judgment, para.228.
[31] See Tadić Appeals Judgment, para. 227-228, and Prosecutor v. Brđanin, Trial Judgment, para. 364.
[32] See Article 28, Rome Statute; see also Prosecutor v. Bemba, ICC-01/05-01/08, Trial Judgment, 21 March 2016, paras. 170-171.

the subordinate.[33] The paradox here is: if the commander can effectively control the autonomous weapon as a fictive "subordinate," it implies that there is no possibility of the autonomous weapon breaching the operator's settings to execute an unlawful attack. Furthermore, in existing mechanisms, a commander's omission manifests as a failure to prevent crimes by subordinates and a failure to punish crimes already or currently being committed by subordinates. Both prevention and punishment are premised on the counterpart being human.[34] Particularly regarding punitive measures, it is hard to imagine LAWS being subjected to human criminal punishment as a "subordinate" – for example, turning off a machine does not equate to a criminal penalty.

D. The Inapplicability of Possible Dogmatic Defenses

Could highly autonomous AI weapon systems be downgraded conceptually, treating the unlawful consequences caused by their spillover as a "weapon malfunction" or stylizing them as a "mistake of fact," thereby subsuming them within the law's traditional interpretative framework? The answer to both questions is negative, for the following reasons.

First, agency spillover is by no means a traditional weapon malfunction or mechanical failure. In the law of armed conflict, traditional weapons are deterministic systems. If errors such as a bomb deviating from its trajectory or a fuze failing occur, they are regarded as accidental events caused by the laws of physics or material fatigue. As long as the weapon itself is lawful and the commander has fulfilled the duty of reasonable care prior to launch, such pure physical failures are typically treated as accidents on the battlefield and are generally not considered acts violating IHL that give rise to State responsibility or individual criminal responsibility.[35]

Equating the agency spillover of LAWS with a weapon malfunction ignores the ontological characteristics of machine learning technology. An ASE is the optimal solution autonomously derived by the system based on the probabilistic calculation of environmental data by its built-in algorithms, while its hardware is functioning normally and its sensors have no faults. For example, if LAWS, exceeding the rules of engagement set by human operators, identify a civilian

---

[33] See Prosecutor v. Bemba, ICC-01/05-01/08, Trial Judgment, paras. 183-184; Prosecutor v. Delalić et al. (Čelebići), Trial Judgment, para. 378.

[34] See generally Kai Ambos, "Superior Responsibility", in Antonio Cassese *et al.* (*eds*), *The Rome Statute of the International Criminal Court: A Commentary*, Oxford University Press, Oxford, 2002, pp. 849-850.

[35] It is worth noting that in relevant jurisprudence of international criminal law, the issue of the margin of error of weapons has been discussed — that is, within what margin an attack's error falls within the tolerance of the attacking party's pursuit of military advantage, and beyond which the attack error becomes an unlawful bombardment, thereby constituting evidence of a suspected war crime. See Prosecutor v. Gotovina et al., IT-06-90-A, Trial Judgments, para.1898; Prosecutor v. Gotovina et al., IT-06-90-A, Appeals Judgments, para.58.

communication tower of a specific structure as an enemy radar and destroy it, it is not because the system is malfunctioning. Rather, its algorithmic agent exercised its endowed functional agency when executing feature matching.

Second, agency spillover should not be simply stylized as a combatant's "mistake of fact." Article 32 of the Rome Statute established "mistake of fact" as a ground for excluding criminal responsibility. Its original intent was to accommodate the cognitive and psychological limitations of humans under the fog of war, such as misjudgements caused by obstructed vision or extreme fear. This is recognised as a lawful justification that negates the subjective criminal intent. Attempting to stylize the identification deviations of LAWS as a machine's mistake is a quintessential anthropomorphizing fallacy. The premise of a mistake is that the subject possesses consciousness and intentionality, capable of generating a misalignment between subjective cognition and objective reality. However, as an "agent without intention,"[36] LAWS do not experience fear, nor are they psychologically affected by battlefield pressure.

Concurrently, the individual commander cannot use this as a defense either. In an ASE, the commander does not directly observe the specific target of engagement. Instead, the cognitive and decision-making processes are substantively outsourced to the machine. Here what the commander faces is not a mistake of fact regarding a specific battlefield reality, but rather a blind reliance on the capability boundaries of the proxy system.

III. The Analytical Perspective of Human-Machine Hybrid Agency

The preceding dogmatic analysis demonstrates how traditional international law falls into a rupture of attribution and evaluation when confronting LAWS in ASE. One of the reasons is that its underlying jurisprudence has experienced a dislocation in both the cognition and handling of a novel type of agency.[37] Since the agency spillover of machines is the direct cause of the dilemma in legal application, the deep excavation and reconstruction of agency as a core jurisprudential concept may serve as a feasible path to break this deadlock.

A. The Evolution of Agency in Legal History

In orthodox legal and moral philosophy, agency has long been deeply bound to free will, moral perception, and intention, being regarded as the exclusive privilege of

[36] See Ian Ayres and Jack M. Balkin, "The Law of AI is the Law of Risky Agents Without Intentions", *University of Chicago Law Review Online*, 2024, available at: https://lawreview.uchicago.edu/online-archive/law-ai-law-risky-agents-without-intentions.

[37] Whether it is IHL, ARSIWA, or ICL, the construction of their legal functions is invariably built upon the single closed loop of the human moral agency.

human beings, or fictive legal persons composed of humans.[38] However, a concise review of legal evolutionary history reveals that the legal system has not exclusively possessed the capacity to handle human agency from beginning to end. Faced with historical impacts of agency originating from heterogeneous sources that exceeded human micro-control, various sub-branches of law have successfully iterated their analytical frameworks on multiple occasions.

One classic example is biological agency. Among the earlier non-human agentic entities addressed in legal history were animals. When a warhorse was startled or a vicious dog injured someone, the animal exhibited a behavioural spillover that deviated from its owner's will. The law never consequently endowed animals with legal personhood or moral subject status, nor did it simply treat them as a mechanical malfunction akin to a broken carriage axle. Instead, by objectively recognizing the animal's instinctive agency, tort law shifted the focus of attribution from seeking the owner's micro-operational fault to controlling the activation of the source of danger. This logic of addressing non-human agency through risk allocation and strict liability is a classic paradigm for the law to handle systems that escape micro-control.[39]

Artefactual agency, or group agency, serves as another example.[40] With the rise of the Industrial Revolution and modern commerce, collective entities such as corporations and organizations demonstrated an agency surpassing that of any single natural person. Through doctrines like vicarious liability, the law recognised that a principal must bear the costs for the autonomous acts of an agent within the scope of authorization. Its jurisprudential essence is: when you expand your own power and pursue your own interests by authorizing an entity with independent capacity for action, you must bear the responsibility for the structural spillover risks it generates.

The law's accommodation of agency is not absolutely premised on the entity possessing human consciousness. When a novel type of agency disrupts the existing

---

[38] See, e.g., Michael Bratman, *Intention, Plans, and Practical Reason*, Harvard University Press, Cambridge, MA, 1987. For the classical link between moral agency and free will, see Immanuel Kant, *Groundwork of the Metaphysics of Morals* (1785), 4:446-448. On the legal tradition's reliance on this presumption, see John Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge University Press, Cambridge, 1998, Ch. 1.

[39] For the Roman law origins, see Justinian, *Digest*, Book 9, Title 1 (*De pauperie*), establishing the *actio de pauperie* for harm caused by animals acting *contra naturam sui generis*. For a comparative analysis of strict liability regimes for dangerous animals across legal systems, see Cees van Dam, *European Tort Law*, 2nd ed., Oxford University Press, Oxford, 2013, pp. 371-380. On the structural parallel between animal liability and emerging AI liability frameworks, see Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, Cambridge, 2021, pp. 158-162.

[40] In legal philosophy, a corporation is viewed as a typical form of Group Agency. The law resolves its spillover effects by granting it derivative rights and imposing vicarious liability. This jurisprudential logic of imposing obligations on non-human entities can be utilized to inspire the regulation of AI. See Christian List, "Group Agency and Artificial Intelligence", *Philosophy & Technology*, Vol. 34, 2021, pp.1213-1242.

order, the corresponding law successfully achieves a re-anchoring of complex causal chains by separating factual capability of action from subjective moral intent and timely updating principles of attribution, with risk liability and vicarious liability being prime examples.

B. The Human-Machine Hybrid Agency

AI has brought about a new form of agency.[41] Drawing on recent research in the philosophy of information and AI law, AI agency can be defined as a factual/functional agency that is independent of human consciousness and moral intent, capable of independently evaluating inputs based on underlying machine learning architectures in open, unstructured physical environments, and autonomously generating intervention strategies and physical consequences.[42] This definition is equally applicable to characterizing the AI-based agency of LAWS.

At this juncture, the law can no longer simply treat LAWS as passive tools or objects, nor can it analyse complex engagements solely through the narrow lens of human subject agency. We must consciously accomplish an ontological perspective shift in legal analysis. On the modern battlefield where LAWS are deployed, focusing solely on singular human agency or pure machine agency is one-sided; the complete analytical object faced by the law is a "Hybrid Agency." It is interwoven from the human (moral) agency embedding the original intent of engagement and initial activation settings, and the functional agency manifested through the machine's subsequent autonomous execution.

To demonstrate why hybrid agency is the necessary conceptual adjustment, we can see how each of the three competing framings produces a specific analytical failure. The first treats LAWS as a **pure tool**, i.e. the human is the sole agent, the weapon is an instrument. In the spillover scenario, this forces the law to evaluate the human's conduct at the moment of the strike, where the human factually had no cognitive access to what happened. We must either attribute to the human a mental state they did not possess, thus distorting *mens rea*; or conclude that no agent is responsible, thus creating an accountability gap. The second treats LAWS as an **autonomous agent** with independent responsibility. This fails not only because LAWS lack legal personhood, but because the machine's agency is constituted by upstream human decisions during design, training, and deployment parameters setting. Treating it as independent severs this constitutive relationship, making human decisions legally invisible. The third and most sophisticated alternative treats the engagement as **sequential agency**, i.e. human acts first, machine acts second, each analysed separately. This fails because it creates a fatal attribution gap at the

[41] L. Floridi, above note 4, p.30.

[42] Floridi explicitly proposes the "Multiple Realisability of Agency" thesis, advocating for defining AI as an "Agency without Intelligence" that requires no human intelligence, intent, or mental state, and arguing that the law should regulate its "functional agency" as an entity agent. See L. Floridi, above note 4, p.30.

seam. The human's agency has ended after they authorized deployment, but the machine's harmful act has not yet occurred, and the causal chain passes through a void attributable to neither. Moreover, IHL evaluates "attacks" holistically; the proportionality principle requires balancing with respect to the whole operation, not isolated segments. Hybrid agency avoids all three failures. It is not a metaphysical claim but a *legal-analytical* claim: when a human delegates cognitive and executive functions to a machine with autonomous behavioural capacity, the legally relevant agent is the composite system.

A boundary question arises: if the commander and the machine form a hybrid, why not extend the concept to include the system's developer, trainer, or procurement officer? The answer lies in operational authorship. Hybrid agency encompasses those actors whose cognitive and executive contributions are functionally indivisible at the level of the specific operational event being evaluated. The commander is included because they authorise the specific deployment in the specific operational context. The machine is included because it executes the tactical decisions within that operational envelope. Upstream contributors such as developers and trainers shaped the system's general capabilities, but they did not author this operation. Their potential liability runs through separate legal channels possibly like product liability or contractual liability, not through the hybrid agent concept.

Introducing the perspective of hybrid agency to analyse the legal issues of LAWS bears significant jurisprudential meaning. It inspires international law scholars examining the legal issues of ASE to stop fruitlessly searching for specific human malice within the machine's black box, or forcefully using "mistake of fact" to cover up the machine's computational autonomy. The human commander provides the normative framework, initial target settings, and authorization for the engagement, while the AI agent fills in the specific spatiotemporal trajectory and target selection. Internally, the agency of these two stages is relatively independent, yet there exist complex, difficult-to-explain or even incomprehensible linkages. However, for external analysis, they exhibit a hybrid agency that can be addressed holistically. Therefore, the determination of whether IHL obligations have been violated can no longer divisibly evaluate the human's button-pressing action in the rear or the machine's physical destruction at the front; instead, the "human-machine hybrid" must be systematically evaluated as an indivisible cognitive and operational unit.

C. The Principal-Agent Analytical Lens and Its Derivative Tools

The essence of hybrid agency is that humans, in order to acquire computational or reaction efficiency beyond their own limits, voluntarily delegate certain cognitive and executive functions to machines. This structural delegation of power and division of function exhibits features familiar from the Principal-Agent (P-A)

relationship in economics and organizational behaviour,[43] making it a valuable analytical lens for illuminating the internal structure of the human-machine hybrid agency.[44] The P-A lens possesses a unique efficacy in addressing the agency conundrum: it can capture the information asymmetry and alignment problem that inevitably accompany agency relationships. In the algorithmic context, the human principal, constrained by cognitive limits, cannot penetrate the algorithmic black box in real time; meanwhile, the machine, acting as the agent, lacks moral intuition and merely pursues the mechanical maximization of its reward function. When environmental data spills over the preset distribution, a phenomenon termed "out-of-distribution" (OOD) in machine learning,[45] the factual agency exhibited by the machine is highly prone to deviate from the human's initial intent. This deviation is not a mere mechanical tool malfunction, but an agency cost inevitably endogenous to the principal-agent structure. Thus, relying on the principal-agent lens allows for the introduction of a mature set of conceptual tools and theoretical logic regarding power delegation, information asymmetry, and the assumption of agency risks and liabilities for subsequent analysis.

However, as an analytical lens, the principal-agent perspective must be combined with domain-specific legal reasoning before it can resolve concrete legal issues. This is because when different legal normative systems confront the same agentic entity, the core questions they pursue and their attribution logics are starkly different. Just as in traditional domestic law, when facing the agency structure of a corporate legal person, contract law tends to apply apparent agency to protect the appearance of transactions, whereas tort law tends to apply vicarious liability to allocate spillover risks. Similarly, facing a human-machine hybrid agency network, we must derive specific analytical tools based on different legal scrutiny objectives.

The first is the Joint Cognitive System model developed based on cognitive science,[46] applicable to evaluating the legality of the behavioural process and decision-making chain. In a highly automated agency network, the high-frequency data processing of machines constitutes a filter for human perception, and humans often lose substantive interventional capacity at the tactical edge. This model posits

---

[43] The Principal-Agent theory in economics and common law has proven to be an effective tool for analyzing the AI black box and the alignment problem. Because AI lacks intrinsic "single-minded loyalty", hidden information and hidden action lead to endogenous agency costs, which naturally aligns in law with the applicative logic of *Respondeat Superior* (vicarious liability). See Anat Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy", *Mitchell Hamline Law Review*, Vol. 46, Iss. 5, 2020, pp.1043-1071. Also see Noam Kolt, *Governing AI Agents*, arXiv preprint arXiv:2501.07913, 2025, pp.18-21.

[44] The P-A lens is employed here not as a full theoretical import but as a structural analogy: in the classical framework, the agent has private interests that diverge from the principal's, producing agency costs. LAWS lack private interests in this sense; the divergence between the machine's behaviour and the human's intent is epistemic rather than motivational.

[45] See Andrew J. Lohn, *Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing out-of-distribution Performance*, arXiv preprint arXiv:2009.00802, 2020, pp.5-6.

[46] See Luciano Floridi, "Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions", *Philosophical Transactions of the Royal Society*, Vol. 374, 2016.

that, in such scenarios, AI is no longer a passive executor independent of the human, but an extension of human cognitive organs; the true bearer of obligations should be defined as the human-machine joint cognitive system. The rationality of this model lies in its establishment of distributed causality. Since the edge decision-making is already indivisible, the law's scrutiny of compliance and feasibility must be shifted leftwards on the timeline, distributed across the entire lifecycle stages of the system, including design, training, and the establishment of macro-constraints.

The second is the Risk Allocation model, which absorbs the jurisprudence of traditional strict liability, applicable to the objective attribution of harmful consequences and loss compensation.[47] Within the agency structure, the factual agency of the agent often produces spillover consequences exceeding the principal's specific instructions. To prevent the principal from exploiting this agency spillover to sever the causal link and evade responsibility, this model analogizes systems with algorithmic autonomy to dangerous animals possessing biological instincts. The logical self-consistency of this tool is built upon the principles of non-reciprocal risk and the alignment of benefit and risk, that is benefit-risk equivalence. Since the principal, for their own benefit, releases an agentic entity with potential destructive power, the incidental damage caused by this entity based on its algorithmic instincts constitutes an inherent internal risk of the system. This model provides theoretical support for establishing absolute vicarious liability without fault and stripping the principal of any room for plausible deniability.

The third is the Reckless Delegation model,[48] focusing on the authorization decision itself, applicable to assessing the individual subjective culpability and punitive evaluation. Noting the legal principle of *nullum crimen sine lege* (no crime without law), this model does not simply advocate mimicking the principal-agent liability models of civil and commercial law to impose the consequences of suspected war crimes from LAWS' agency spillover engagements upon human subjects. Rather, it argues for shifting the anchor of evaluation from the transgressive acts at the agent's behavioural end to the principal's frontend authorization and decision-making behaviour. This tool demonstrates that if a principal, fully aware that the introduced black-box agent poses a systemic risk of misalignment in complex environments, still delegates lethal force to it without physical limitations or architectural constraints, this decision itself embodies an extreme reckless disregard for the lawful rights of others. This interpretative logic can convert the direct intent

---

[47] The risk allocation approach to AI liability draws on the tradition of strict liability for ultrahazardous activities. See Anat Lior, "AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy", *Mitchell Hamline Law Review*, Vol. 46, Iss. 5, 2020, pp.1043-1071; see also Gerhard Wagner, "Robot Liability", in Sebastian Lohsse *et al. (eds)*, *Liability for Artificial Intelligence and the Internet of Things*, Nomos, 2019, pp. 27-62.

[48] On the concept of reckless risk-creation as a basis for criminal liability in the context of autonomous systems, see I. Ayres and J. Balkin, above note 36. For the broader doctrinal foundations of shifting *actus reus* to the deployment decision, see Noam Kolt, Governing AI Agents, arXiv preprint arXiv:2501.07913, 2025, pp. 18-21.

for specific harmful outcomes into recklessness in creating an impermissible agency risk, thereby providing a viable legal reasoning path for punishing irresponsible algorithmic applications while firmly upholding the baseline of subjective attribution.

To avoid collapsing into *de facto* strict criminal liability, this model incorporates a culpability threshold. The delegation is reckless when the commander knew or willfully disregarded a substantial and unjustifiable risk of agency spillover in the specific operational context *and* failed to implement available architectural constraints that would have materially reduced that risk. If a commander implemented every reasonably available safeguard, such as the adversarial testing, geofencing, target-class restrictions, override mechanisms, and still experienced spillover, he or she has not engaged in reckless delegation. That is treated as a reasonable decision with a bad outcome, not a culpable one.

IV. Transcending the Legal Dilemmas in Agency Spillover Engagements

The perspective of human-machine hybrid agency and the principal-agent analytical lens provide the possibility to reactivate existing law. The fundamental utility of this theoretical perspective is that it elevates the analytical unit of legal evaluation from isolated tactical actions to a systemic network of authorization and decision-making. It reveals that the agency spillover of machines is not an incomprehensible physical malfunction, but an endogenous, structural agency cost inevitably arising after the delegation of power. Based on this reconceptualisation, we need not search in vain for human micro-faults within the machine's algorithmic black box, nor do we need to endow the machine with a fictive legal personhood. Instead, we can utilize specific analytical tools to attempt to reshape the connotations of old legal concepts through evolutionary and teleological interpretation, without departing from the existing positive law (*lex lata*) framework.[49] In doing so, we could achieve a substantive transcendence of the three major legal dilemmas within the realm of legal hermeneutics.

A. Dogmatic Updating of the Application of Substantive IHL Rules

The primary IHL dilemma revealed previously is that the high-frequency computational agency of machines deprives humans of the physical possibility of intervention at the tactical end, resulting in the failure of both the feasibility requirement of the precautionary principle and the subjective balancing required by the principles of distinction and proportionality. The Joint Cognitive System analytical tool may prove helpful to transcend this dilemma.

[49] A methodological note is warranted here. This paper's commitment to *lex lata* is not merely strategic but substantive. The primary obligation of legal scholarship confronting a new technological reality is to test rigorously how far existing law can stretch through interpretation before concluding that new law is needed.

From the perspective of legal hermeneutics, API Article 57 requires the attacker to take "all feasible precautions" to verify targets and control incidental damage. In the context of traditional weapons, the literal meaning of "feasible" usually refers to visual confirmation or radar calibration at the very moment of engagement. However, in a human-machine hybrid agency network, because the terminal decision has been substantively delegated to the agent, demanding humans to fulfill verification and suspension obligations in millisecond-level engagements is physically no longer "feasible." The ultimate purpose of IHL is to protect civilians and limit unnecessary suffering. To prevent this normative purpose from being circumvented by technology, the conceptual extensions of feasibility and attack must be dynamically expanded. Based on the distributed causality established by the Joint Cognitive System, since human cognition and machine algorithms are indivisible during the execution phase, the law's scrutiny of the obligations of distinction and proportionality must undergo a systemic "shift left" – i.e. to expand the compliance span to *ex ante* lifecycle – on the timeline.

This means that feasible precautions under interpretative legal theory are no longer confined to the moment the trigger is pulled but are lawfully expanded to the entire lifecycle of the system.[50] For example: Did the commander and operator conduct sufficient adversarial testing (red-teaming) on out-of-distribution data for the specific theater of operations prior to deployment? Did they establish strict geofencing and forced-sleep time windows for the system in accordance with the ROE? Only if the commander has fulfilled the compliance obligations of the joint system at the frontend stages of system constraint setting and deployment authorization can they be deemed to have fulfilled the substantive obligations of IHL. Conversely, neglecting to establish architectural constraints at the frontend, thereby leading to agency spillover, would constitute a suspected violation of the principles of distinction or proportionality.

B. Evolutionary Interpretation of the Attribution Rules in State Responsibility

In the *Namibia* and *Costa Rica v. Nicaragua* cases,[51] the International Court of Justice (ICJ) established that generic terms with evolutionary meanings should be interpreted in accordance with the legal and factual realities at the time of

[50] Recent international practice emphasizes that human agency and IHL compliance cannot be evaluated solely at the moment of engagement, but must permeate the entire lifecycle of AI (from research and development, testing, and procurement to deployment and post-use assessment). See The AutoPractices Team (Ingvild Bode *et al.*), *Strengthening Human Agency in the Military Domain* (Center for War Studies, 2026), pp. 21-27. See also the CCW GGE Rolling Text on LAWS, 18 December 2025, CCW/GGE/LAWS, especially Box IV on human responsibility across the lifecycle; and also Jonathan Kwik, "Iterative Assessment for Military Artificial Intelligence (AI) Systems", in Bérénice Boutin, Taylor Kate Woodcock, Sadjad Soltanzadeh (*eds*) *Legal, Ethical, and Technical Dilemmas in Military Artificial Intelligence*. T.M.C. Asser Press, The Hague, 2026.

[51] ICJ, *Legal Consequences for States of the Continued Presence of South Africa in Namibia (South West Africa)*, Advisory Opinion, ICJ Reports 1971, p. 31, para. 53; ICJ, *Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua)*, Judgment, ICJ Reports 2009, p. 213, para. 64.

application. By introducing the analytical perspective of hybrid agency, when conducting legal attribution arguments under State responsibility law for suspected unlawful acts arising from LAWS in ASE, the following interpretative updates can be made; and it is useful to present them in order of doctrinal conservatism.

The most conservative textual anchor is Article 7 of ARSIWA, which provides that the conduct of an organ of a State shall be considered an act of the State "even if it exceeds its authority or contravenes instructions." The human operator or commander who deploys LAWS is a State organ under Article 4. Under hybrid agency, the human-machine system is the operational unit through which this State organ exercises military force. When the system produces an ASE, the resulting conduct, namely the unauthorized strike, is conduct that *exceeds the authority* given to the system by the human component. Article 7 captures precisely this situation. Without hybrid agency, the State could argue that the organ's own conduct, i.e. pressing the deploy button, was lawful, and the unlawful conduct was the machine's alone. With hybrid agency, the engagement is indivisible: the organ acted through the machine, and the machine's spillover remains part of the organ's *ultra vires* conduct.

Alternatively, by linking the suspected unlawful conduct holistically to the hybrid agency without delving into its internal agency allocation mechanisms, the human-machine hybrid as a whole can be treated as the "person or entity" exercising elements of governmental authority under Article 5 of ARSIWA. There is currently no general framework in international law regulating State responsibility for inanimate objects, with only specific regimes existing, such as those applicable to space objects or transboundary harm caused by hazardous activities.[52] However, as previously noted, an authorized entity can encompass a wide range of forms, provided it is empowered by the internal law of the State. In other words, within the operational margin of treaty interpretation, there is an opportunity to achieve attribution justification through this path.

Lastly, the interpretation of ARSIWA Article 8 can also be navigated. Within the interpretative framework of existing law, the "effective control" standard established by the ICJ in classic jurisprudence typically requires the State to exercise micro-command over specific tactical operations. Confronted with LAWS, we could apply a teleological interpretation to theoretically upgrade the concept of "control" in Article 8. Since the State, acting as the principal through its commanders and operators, voluntarily delegates lethal force to a machine agent to gain an asymmetric military advantage, this macro "systemic authorization" inherently constitutes a form of "control" consistent with the normative intent of ARSIWA. The spillover attacks emerging from the machine based on its reward function as agency costs belong to the endogenous risks within the inherent probabilities of this hazardous

[52] See Magdalena Pacholska, "Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective", *Israel Law Review*, 2023, pp.20-21.

system. Therefore, interpreting such acts as "conduct under the direction or control of a State" aligns with the normative intent of ARSIWA, thereby blocking the State's attempt to use the technological black box to sever the causal chain.

C. Reinterpreting the Subjective Elements of Individual Accountability in ICL

In the realm of ICL, the pursuit of individual criminal responsibility is obstructed by the stringent subjective requirements of Article 30 of the Rome Statute, namely, the commander's lack of specific intent and knowledge regarding the particular civilian casualties caused by the machine's spillover. Utilizing the Reckless Delegation model, we may complete the closed loop of criminal dogmatics.[53] Simply put, in positive law, we do not attempt to prove that the commander had direct intent toward the specific civilian ultimately bombed by mistake, which contradicts the facts or presents legal dilemmas; instead, we propose treating the commander's deployment act as the *actus reus* under scrutiny.

Utilizing the Reckless Delegation model, the anchor of evaluation shifts from the unlawful moment of engagement to the authorization decision. When a commander knows full well that a certain type of LAWS, acting as a black-box agent, poses an extremely high risk of misalignment and spillover in a complex electromagnetic urban theater, yet still signs the deployment order without implementing effective physical geofences in pursuit of tactical advantage, then, this decision arguably satisfies Rome Statute Article 30's "knowledge" requirement. Article 30(2)(b) defines knowledge as awareness that "a consequence will occur in the ordinary course of events." A commander who deploys a system with a known risk profile of autonomous misidentification in the given operational context is aware that civilian harm is a statistically predictable consequence of that deployment. In other words, it may not occur in every instance, but rather "in the ordinary course of events" across deployments of this type in similar conditions.

Should the above interpretation be considered too expansive, an alternative basis exists. Article 30 opens with the caveat "unless otherwise provided," acknowledging that specific crimes may carry different *mens rea* requirements under other sources of law. Customary international law, as reflected in the practice of international criminal tribunals including the ICTY, recognises that *dolus eventualis* or recklessness may satisfy the *mens rea* for certain war crimes.[54] Where this is the case, the recklessness standard, which the deployment decision described above clearly meets, applies independently of Article 30's default threshold.

---

[53] As identified in Section II.C, even where recklessness is accepted as the applicable mental state, two structural problems must be overcome: the *actus reus* must be reconceived, and the causal link must be established through the algorithmic black box. The Reckless Delegation model addresses both.

[54] Jurisprudentially, regarding "risky agents without intentions," the law may abandon the search for the underlying micro-level subjective intent and instead determine the recklessness or gross negligence of the deployer through objective standards of risk control. See I. Ayres and J. Balkin, above note 36.

A further challenge is that criminal law requires specificity in the risk the defendant foresees. In other words, a vague sense that "something might go wrong" is insufficient to establish *dolus eventualis*.[55] The hybrid agency framework helps address this in two ways. First, the Joint Cognitive System model specifies *what* the commander should know. The system's documented failure modes, its performance profile in adversarial testing, and the gap between training distribution and operational environment may all be included. This is not diffuse unease but technically specific risk awareness. Second, the Reckless Delegation model defines *what kind of risk* counts: agency spillover producing civilian harm from autonomous misidentification in the specific type of operational environment where the system is deployed. The commander's awareness is assessed against the system's known risk profile, which is a concrete, documentable set of facts, rather than against a general anxiety about AI unpredictability. That said, borderline cases will inevitably arise, and the precise threshold of risk specificity will require judicial development.

V. Conclusion

The rapid evolution of military artificial intelligence has pushed IHL to an unavoidable ontological crossroads. The ASE declares the localized failure of traditional legal application built upon the presumption of a pure human moral subject. The key to breaking this structural crisis is neither mechanically reducing the algorithm to a mere tool malfunction nor taking a radical leap to endow it with personhood, but rather facing head-on the human-machine hybrid agency in modern armed conflict.

By employing the Principal-Agent relationship as an analytical lens and developing three derived analytical tools, this paper demonstrates that *Lex Lata* still possesses robust resilience to regulate algorithmic violence. Through the Joint Cognitive System, IHL compliance obligations are anchored to the system's lifecycle compliance. Through the Risk Allocation model, State attribution in ARSIWA is extended to the assumption of systemic agency costs. Through the Reckless Delegation model, the punitive focus of ICL can be redirected to the decision-making hub that abuses black-box algorithms.

This series of jurisprudential reconstructions directly responds to and empowers the ongoing governance processes of global military AI. It provides a theoretical basis for implementing meaningful human control in the Geneva CCW GGE negotiations,[56] implying that the failure of micro-control must be replaced by

---

[55] See Jonathan Kwik, "The Conceptual Roots of The Criminal Responsibility Gap in Autonomous Weapon Systems", *Melbourne Journal of International Law*, Vol. 24, 2023, pp. 20-21.

[56] See Chair's summary – Second 2025 session of the GGE on LAWS - 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, 20 October 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_ Weapons_-

architectural red lines. Concurrently, it injects a hard-law accountability logic into the lifecycle governance advocated by the GC REAIM initiative.[57]

Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_ Systems_(2025)/CCW-GGE.1-2025-WP.9_-_Chair's_summary.pdf

[57] See GC REAIM, "*Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain*", September 2025, available at: https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf

Digital Forensics As A Verification Method For AI-Powered Autonomous Weapons Systems

Dexter "Drexx" Laggui[*]

ABSTRACT

States can declare restraint in using AI-enabled military systems, but at present they can neither prove this nor verify each other's claims. This article argues that the accountability and compliance gaps in autonomous weapons systems are, at root, verifiability gaps. Drawing on digital forensics standards (ISO/IEC 27001, 27037, 27042, 27043 and 42001) and electronic-evidence practice, it proposes forensic readiness as the inspection methodology: the designed-in logging, retention and authentication of mission, control-plane, training and runtime records. The resulting posture, which this article calls Evidentiary Custodianship, makes command responsibility, weapons review and meaningful human control testable rather than declared.

Keywords: autonomous weapons systems; digital forensics; forensic readiness; Evidentiary Custodianship; verification; arms control; meaningful human control; appropriate levels of human judgment; command responsibility; Article 36 weapons review; ISO/IEC 27037; ISO/IEC 27043; ISO/IEC 42001; electronic evidence; international humanitarian law

1. Introduction: When Restraint Cannot Be Verified

Imagine two States that have agreed, in principle, that lethal autonomous targeting in densely populated areas requires meaningful human authorization for each engagement. Both deploy AI-enabled targeting systems. Both publish declarations of compliance.

After a strike that kills civilians, each accuses the other of breaching the agreement. Each insists its own system kept a human in the decision loop. Neither can produce records that show where its approval gate sat at the moment of the strike, what its operator saw, what the system's confidence output was, what time elapsed between classification and actuation, or whether known failure modes had been documented before the system was fielded. The investigation stalls not because the law is silent, but because the systems are.

---

[*] Dexter "Drexx" Laggui is the Principal Consultant at DREAMLab OPC and a digital forensics expert and an experienced expert witness. He was the first chairperson of the Ad Hoc AI Committee of the US Scientific Working Group on Digital Evidence (SWGDE), contributed to ISO/IEC JTC 1/SC 42 on AI trustworthiness for the Philippine delegation, and has undertaken advisory missions for the UNODC and UNIDIR.

This is the AI-armed-conflict variant of an old structural problem. Defensive design can be indistinguishable from offensive design at a distance.[1] Ambiguous compliance can pass for either restraint or cheating. Accelerated development becomes the rational response to uncertainty. The result is a security dilemma that no amount of declaratory diplomacy can resolve.[2] Trust requires inspection, inspection requires evidence, and evidence requires that systems be designed in advance to produce it.

The strategic logic is summarized in Figure 1. Restraint that cannot be proved is rationally punished. Ambiguous compliance, meaning declared restraint with undeclared development, is rationally rewarded. Without a verification methodology, the equilibrium settles into mistrust and mutual capability accumulation, even where both parties would have preferred restraint. The methodology this paper proposes operates first as a unilateral self-binding measure, and it dissolves the mutual form of the dilemma only on reciprocal adoption; the full mutual case is deferred to the verification clause of Section 9.

State A / State B	True restraint	Ambiguous compliance	Accelerated development
True restraint	Stable cooperation; lower escalation risk; strategic parity.	State A is exposed to State B's hidden advantage; norms exist but are fragile.	State A becomes vulnerable to overt dominance.
Ambiguous compliance	State A gains crisis advantage while eroding trust.	A false cooperation equilibrium emerges: norms exist, covert development persists.	State A faces partial disadvantage and growing mistrust.
Accelerated development	State A dominates a constrained opponent.	State A dominates, but pays reputational and escalation costs.	Full arms race; instability; security dilemma.

Figure 1. The AI Prisoner's Dilemma in military restraint. Author-created, after the workshop notes on file with the author.

[1] This is the AI-armed-conflict instantiation of what international-relations scholars call the security dilemma: defensive measures that resemble offensive measures generate counter-measures, even where the original intent was restraint. See Robert Jervis, "Cooperation Under the Security Dilemma", *World Politics*, Vol. 30, No. 2, 1978, pp. 167–214.

[2] International Committee of the Red Cross, *ICRC Position on Autonomous Weapon Systems*, Geneva, 12 May 2021.

The argument of this paper is therefore narrow but, if accepted, consequential. International humanitarian law already supplies the substantive obligations that govern the use of AI-enabled force: the duty to review new means and methods of warfare under Article 36 of Additional Protocol I,[3] the precautionary obligations of Article 57,[4] the principles of distinction and proportionality,[5] the doctrine of command responsibility codified in Article 28 of the Rome Statute,[6] and, for States like the Philippines that have legislated international crimes domestically, the corresponding criminal-law provisions of Republic Act No. 9851.[7] What is missing is not the legal architecture but the evidentiary architecture. Digital forensics, properly anchored in international standards and built into procurement, can supply that architecture. It can convert claims about human control, command knowledge, weapons review and treaty compliance into questions a tribunal or an inspector can actually adjudicate.

This paper does not argue that a new treaty is unnecessary. Other contributions to this special edition, including Zhixiong Huang and Haokun Yu's analysis of why a new instrument is *de jure* necessary,[8] make the legislative case in detail. The argument here is complementary. Any treaty, and any existing rule that already binds States, will be only as strong as the verification capability that can test compliance with it. Without that capability, the *de jure* framework, old or new, risks becoming a regime of declarations and counter-declarations in which the speed of AI-driven incidents outpaces the speed at which evidence can be gathered, authenticated and assessed.

The argument moves from the problem to the method to the law, and then tests itself. Section 2 sets out the verification problem and locates it within current Autonomous Weapons Systems (AWS) discourse. Section 3 explains the AI targeting and decision-support chain for non-technical legal readers, from data ingestion to actuation, and introduces the vocabulary of the control plane and approval gate. Section 4 presents the three-stage evidentiary spine of forensic-ready AI: training, post-training alignment, and inference runtime. Section 5 develops the central proposal of digital forensics, anchored in ISO/IEC 27001, 27037, 27042,

---

[3] Protocol Additional (I) to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977, Art. 36.
[4] *Ibid.*, Art. 57.
[5] *Ibid.*, Arts 48, 51 and 52.
[6] Rome Statute of the International Criminal Court, 2187 UNTS 3, 17 July 1998, Art. 28.
[7] Republic Act No. 9851, *Philippine Act on Crimes Against International Humanitarian Law, Genocide, and Other Crimes Against Humanity*, 11 December 2009, Sec. 10. The Philippines is also a party to Additional Protocol I (ratified 30 March 2012), so for that jurisdiction the treaty obligations invoked here and the domestic criminal vehicle coincide. The Philippines withdrew from the Rome Statute effective 17 March 2019; this paper invokes Article 28 only as doctrinal reference, with Republic Act No. 9851 as the operative domestic vehicle, and does not rely on the Court's jurisdiction over the Philippines.
[8] Zhixiong Huang and Haokun Yu, "*De Jure* Necessity vs *De Facto* Improbability? Establishing a New Treaty for Artificial Intelligence in the Military Domain", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.129-159.

27043 and 42001, operationalized through procurement, as the verification methodology, and introduces the framework of Evidentiary Custodianship. Section 6 anchors the proposal in existing doctrine, including meaningful human control, appropriate levels of human judgment, command responsibility, Article 36 review and Article 57 precautions. Section 7 stress-tests the framework against four case studies, including learned deception. Section 8 surveys implementation barriers (slow mutual legal assistance, judicial education, infrastructure constraints, the automated-forensics paradox and standards convergence) and addresses the fairness point that forensic readiness can also exonerate honest commanders. Section 9 translates the argument into the minimum content of a future verification clause. Section 10 concludes.

2. The Verification Problem

2.1. What current scholarship does, and does not, supply

A great deal has been written about why AI-enabled military systems should be regulated and how responsibility for their effects should be allocated. International law already applies to new means and methods of warfare. The International Court of Justice held in *Legality of the Threat or Use of Nuclear Weapons* that the established principles and rules of humanitarian law apply to weapons invented after the rules were drafted.[9] State delegations and the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts have proceeded on the same premise: the use of force by autonomous and AI-enabled systems remains within the existing legal framework, and human responsibility must be retained.[10] Civil society and several State delegations have pressed for "meaningful human control" as a normative threshold.[11] Others, principally the United States, have argued for "appropriate levels of human judgment over the use of force" as the operative standard.[12] Scholars have proposed expanding command responsibility,[13] introducing dedicated war-tort regimes,[14] and treating AWS as inherently incapable

[9] International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, ICJ Reports 1996, p. 226, paras 86–87 (holding that the established principles and rules of humanitarian law apply to all forms of warfare and to all kinds of weapons, including those of the future).

[10] Convention on Certain Conventional Weapons, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, *Guiding Principles* affirmed by the High Contracting Parties, UN Doc. CCW/MSP/2019/9, 13 December 2019, Annex III, esp. principles (a) and (d).

[11] ICRC, above note 2; see also Article 36 (NGO), *Key Areas for Debate on Autonomous Weapons Systems*, CCW submissions and commentary, 2014–2024.

[12] US Department of Defense, *Directive 3000.09: Autonomy in Weapon Systems*, originally issued 21 November 2012, revised 25 January 2023.

[13] For one influential treatment, see Nehal Bhuta *et al.* (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, Cambridge, 2016.

[14] Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons", *University of Pennsylvania Law Review*, Vol. 164, 2016, p. 1347.

of compliance with proportionality and distinction.[15] These contributions have done important work. They have not, however, supplied a methodology for testing whether any particular deployment, in any particular incident, complied with whichever standard a State purports to have observed.

The literature on AI explainability and traceability is closer to the verification problem, but most of it is framed around regulatory disclosure to civil oversight bodies rather than adversarial inspection by another State or a tribunal.[16] Arms-control verification literature, by contrast, has matured around physical assets such as fissile material, warhead counts and missile launchers, and has only recently begun to grapple with software-defined systems.[17] The result is a structural gap. The AWS literature does not look like arms-control inspection, and arms-control inspection literature does not look like software audit.

Bridging that gap is this paper's contribution. The literature on digital forensics and on electronic-evidence admissibility are technically mature and legally tested in domestic settings, including in the Philippines through the Rules on Electronic Evidence.[18] What has not happened is their adaptation for the verification of State compliance with international rules governing AI-enabled force.

2.2. Why declarations fail in software-defined systems

Three features make AI compliance especially difficult to observe. First, capability can change without obvious physical change. A conventional munition is tied to visible engineering. An AI-enabled system can change through retraining, fine-tuning, data integration, threshold adjustment or a sensor-fusion update. A system inspected in January may not be the same system deployed in March, even if the airframe is unchanged. Second, the legally decisive fact is often relational rather than material. The question is not only "what is the weapon" but "how was human authority structured around it," and the same physical platform may preserve meaningful human control under one mission configuration and lose it under another. Third, compliance depends on records that are easy to omit, overwrite or classify. Training provenance, evaluation reports, model-card history, operator-interface logs and inference logs are volatile. If no preservation duty exists before deployment, post-incident investigators may find the evidence already gone.

---

[15] Human Rights Watch and International Human Rights Clinic, *Losing Humanity: The Case Against Killer Robots*, Cambridge MA, November 2012.

[16] For a survey of the explainability literature in legal contexts, see Brent Mittelstadt, Chris Russell and Sandra Wachter, "Explaining Explanations in AI", *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 279–288.

[17] United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches*, UNIDIR Resources No. 6, Geneva, 2017; for the limits of physical-asset analogues, see also International Atomic Energy Agency, *IAEA Safeguards Glossary*, 2022 ed., Vienna, 2022 (noting that safeguards are organized around nuclear material and facilities).

[18] Philippine Supreme Court, *Rules on Electronic Evidence*, A.M. No. 01-7-01-SC, 17 July 2001.

Three further features make declarations especially fragile. Definitional uncertainty over what counts as autonomy, what counts as "in the loop" or what counts as a meaningful intervention window leaves room for strategic ambiguity. Update opacity allows a system to move from compliant to non-compliant configurations through changes that look small to a non-technical reviewer but matter operationally. Selective disclosure lets a State publish doctrine and a high-level architecture while withholding the operational records that would test whether doctrine and architecture matched reality.

2.3. Verifiability as a precondition for legal certainty

Huang and Yu argue that legal certainty is an unmet precondition for compliance with rules governing military AI.[19] That argument can be extended. Legal certainty in this context has at least two components: certainty about what the rule requires, and certainty about whether the rule was observed in a given case. Existing scholarship has focused almost entirely on the first. The second is what verification supplies.

A compliance standard that cannot in principle be tested against evidence is, in any operational sense, a slogan rather than a standard. A standard that can be tested but only against evidence the deploying State controls and need not produce is barely better. The closer a verification methodology can come to producing reviewable, authenticated records that bind the deploying State to the facts of its own operations, the closer law moves to a regime under which AI-enabled force can be legally restrained.

3. The Technical Problem for Legal Readers

The vocabulary in this section matters because the legal arguments that follow turn on it. Readers familiar with the technical detail can skim. Subsequent doctrinal arguments about who exercised authority, what records existed, and what humans saw and could do depend on the precision of these terms.

3.1. The decision-support chain

A typical AI-enabled targeting or decision-support pipeline has six stages.[20] Data ingestion takes in sensor outputs (imagery, signals, network telemetry, prior intelligence) and external data feeds. Classification applies a trained machine-learning model to identify or categorize patterns, for example by matching imagery against a learned model of "vehicle of interest" or "hostile actor pattern".

---

[19] Huang and Yu, above note 8, Section 2.5, "Where Does the Necessity Come From: Legal Certainty as a Precondition for Compliance", pp.139-142.

[20] For an introduction to AI decision-support pipelines in military settings, see Arthur Holland Michel, Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making, ICRC, Geneva, 2024.

Recommendation combines classifications with rules of engagement, mission parameters and confidence thresholds to produce an output for a human, or for a downstream system, to act on. Approval, where present, is the step at which a human authorizes the next action. Actuation is the kinetic or non-kinetic action itself: a strike, an alert, a hand-off to another system. Logging records what occurred, in some level of detail, for after-the-fact reconstruction.

Two points matter for legal readers. First, every one of these stages is configurable. A system can be operated with the approval step at the level of each individual strike, at the level of a target list approved earlier, or with no approval at all within a pre-authorized mission envelope. Second, the logging stage is not automatic in any legally meaningful sense. AI systems generate trails only to the extent they are designed and procured to do so. Forensic evidence depends on what the system was built to record, what it was permitted to retain, and what was authenticated against tampering.

3.2. Training, post-training alignment, inference runtime

A second triad is essential. The training stage is where a model learns from data. The data, the filtering choices, the validation set, the documented limitations and the failure modes that were tolerated all shape what the model is capable of and what it is known not to handle reliably.[21] The post-training alignment stage, including supervised fine-tuning, reinforcement learning from human feedback, preference optimization, safety calibration and reward-model design, is where deliberate human choices reshape the model's behavior after basic capability has been established. This is where decisions about how cautious or how aggressive the model will be are encoded.[22] The inference runtime is the actual operational use, the moment a soldier, analyst or commander has a system in front of them and acts on its outputs.

Each of these stages is also a stage at which evidence is produced. Training-stage records (data provenance, validation results, known-failure documentation) speak to what was knowable about the system before deployment. Alignment-stage records (preference data, reward functions, fine-tuning logs, safety-tuning decisions) speak to what was deliberately reinforced or relaxed by identifiable people. Runtime records (inputs, classifications, confidence outputs, operator displays, approval timestamps, override attempts, actuation logs) speak to what happened in the field. Together, the three stages allow the law to ask not only what the system did, but what its operators already knew and had chosen before it did it.

The two sequences are not redundant; they index different doctrinal questions. The operational pipeline answers where authority and action sat at the moment of use, the question that drives meaningful-human-control analysis, real-

[21] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Gaithersburg, January 2023.
[22] For a technical overview of post-training alignment methods, see Yuntao Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback", arXiv:2212.08073, December 2022.

time precaution under Article 57, and the proximate causation inquiries on which command responsibility depends. The lifecycle triad answers what was known, chosen and changed before that moment, the question that drives *mens rea* analysis, weapons review under Article 36, and the foreseeability findings that distinguish accident from negligence. A legal reader needs both: pipeline evidence without lifecycle evidence may show a defective act without showing a culpable mind, and lifecycle evidence without pipeline evidence may show culpable design choices without connecting them to a particular use of force. The methodology in Section 5 keeps both sequences in view because legal accountability requires both.

3.3. The control plane and the approval gate

Two further pieces of vocabulary do most of the legal work. The control plane is the structure of human authority that sits above operational behavior: who is permitted to deploy the system, under what mission parameters, with what oversight mode, with what override authority, and under whose supervision.[23] The control plane governs the system. The system performs the task.

The approval gate is the specific point in the decision-support chain at which the system must pause and obtain human authorization to proceed.[24] The location of the gate is itself a legally consequential design choice. A gate placed at each individual engagement preserves a tight connection between human decision and lethal outcome. A gate placed at mission activation or rules-of-engagement configuration preserves human authority over the framework, but not necessarily over the act. The forensic question is not only whether a gate existed, but where, in time and in authority, it sat.

A further distinction matters. Human-In-The-Loop (HITL) systems cannot proceed without explicit human authorization at the relevant step. Human-On-The-Loop (HOTL) systems execute within authorized parameters while a human supervisor monitors and retains authority to interrupt. Human-Out-Of-Loop (HOOL) systems operate within pre-authorized parameters with no real-time human participation in the relevant operational window.[25] HOTL is the hardest mode to evaluate, because authority that is theoretically available may not be practically exercisable, for example when the time between classification and actuation is

---

[23] The phrase "control plane" is borrowed from network and distributed-systems engineering, where it denotes the layer that decides how data and processes are routed, distinct from the data plane that performs the routing. The metaphor is used here to capture the distinction between governance and operational behavior.

[24] For a related taxonomy in a different automation domain, see SAE International, *SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, revised April 2021.

[25] For the loop-position taxonomy in military AI specifically, see Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies: A Theoretical and Operational Perspective*, T.M.C. Asser Press, The Hague, 2024.

shorter than human cognitive response time, or when the system produces output volumes that exceed human review capacity.[26]

Fan Yang's analysis of human-machine hybrid agency is helpful here precisely because it shifts the legal object of inquiry away from the final machine act and toward the structured network of human delegation and machine operation in which agency is exercised.[27] Forensic readiness, on the argument of this paper, gives that hybrid agency a record. It shows where the human delegated, where the machine operated, where agency spilled beyond expectation, and whether the human system had safeguards for that spillover.

4. The Evidentiary Backbone of Forensic-Ready AI

The familiar phrase "the machine always leaves a trail" needs qualifying. AI systems can in principle be designed to leave detailed legal trails. By default they do not. The phrase is true only in the trivial sense that some artefacts always exist. Legally meaningful trails depend on logging being designed in, retention being governed, authentication being preserved against tampering, and interpretability being supported. ISO/IEC 27037 explicitly addresses identification, collection, acquisition and preservation of digital evidence;[28] ISO/IEC 27042 addresses analysis and interpretation.[29] These standards apply to AI evidence as much as to any other form of digital evidence, but only if the systems were procured to comply with them.

A working distinction at the outset: transparency makes information visible; verification makes a claim testable. A State may be transparent about general principles while remaining unverifiable in operational design. Conversely, a State may legitimately keep some technical details confidential while still preserving enough authenticated evidence for a tribunal, inspector or cleared review body to test a compliance claim. The aim of forensic readiness is disciplined inspectability rather than total exposure: enough record to settle the question, not enough to open every operational secret.

---

[26] For the foundational analysis of cognitive overload in machine-mediated decision-making, see Lisanne Bainbridge, "Ironies of Automation", *Automatica*, Vol. 19, No. 6, 1983, pp. 775–779.

[27] Fan Yang, "Transcending the Dilemma of Applying IHL to LAWS in Agency Spillover Engagements: A Legal Analytical Perspective on Human-Machine Hybrid Agency", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.190-212.

[28] International Organization for Standardization, *ISO/IEC 27037:2012 – Information technology – Security techniques – Guidelines for identification, collection, acquisition and preservation of digital evidence*, Geneva, 2012.

[29] International Organization for Standardization, *ISO/IEC 27042:2015 – Information technology – Security techniques – Guidelines for the analysis and interpretation of digital evidence*, Geneva, 2015.

4.1. Mapping artefacts to legal questions

Table 1 sets out, for each main category of artefact a forensic-ready AI system should preserve, the verification question it answers and the legal anchor it speaks to.

Artefact category	Verification question	Legal anchor
Pre-deployment testing, validation and red-team records	What was known about the system's limits before it was fielded?	Article 36 review; foreseeability; "should have known"
Post-training alignment records (fine-tuning, preference data, reward design)	What trade-offs were deliberately reinforced or relaxed by identifiable people?	Knowledge; conscious disregard; institutional intent
Authenticated mission parameter files	Who authorized what operational envelope, and under what limits?	Effective control; rules of engagement; mission authorization
Approval-gate settings and version history	Where, in the chain, did human authority sit at the time of the act?	Meaningful human control; appropriate levels of human judgment
Command-message logs with identity-linked credentials	Which identifiable person issued which order, when, and under what authority?	Command responsibility; chain of command; AP I Art. 57(2)(a)(i)
Sensor archives covering the engagement window	What inputs did the system actually receive?	Distinction; proportionality; factual reconstruction
Inference records: classifications, confidence outputs	What did the system conclude, and how confident was it?	Distinction; precaution; foreseeability of error
Operator-interface logs	What did the human see, and when did they see it?	Real human judgment; informed authorization; automation-bias analysis
Override records, alert logs, intervention-window data	Was intervention practically possible? Was it attempted?	Constant care; failure to act; *mens rea*
Runtime records linking	What actually happened,	Causation;

inputs, outputs, warnings, human actions, system actions	in what order, under whose authority?	reconstruction; proof of breach
Retention, access and hash/chain-of-custody records	Has the evidence been preserved without alteration?	Admissibility; chain of custody under ISO/IEC 27037 and the concurrent process of ISO/IEC 27043

Table 1. Forensic artefacts, verification questions and legal anchors. Author-created; categories drawn from ISO/IEC 27037 and 27042 conventions.

The table makes two things visible at once. First, almost every doctrinal question (about meaningful human control, command responsibility, weapons review, due care) has at least one corresponding artefact category that, if preserved and authenticated, can render the doctrinal question evidentiary rather than rhetorical. Second, the absence of any of these artefacts is itself an evidentiary fact. It signals what a State or a deploying force was willing to be held accountable for.

One category needs caution. A confidence output is not neutral ground truth; it is a model-internal number, possibly poorly calibrated, produced by the same probabilistic system under examination. The figure is probative of what the system reported, but its weight depends on a separate showing that the confidence measure was validated, with a known error rate, on the kind of input at issue, which is the validation that emerging forensic practice for machine-learning outputs treats as a precondition of admissibility.[30] A forensic-ready regime should preserve that validation record alongside the runtime log.

4.2. The temporal sequence as evidence

Among these artefacts, the temporal sequence (what happened in what order, with what time elapsed between events) is uniquely powerful. If the gap between classification, any required human approval and actuation is 300 milliseconds, the timing record leaves little room for any account of real-time human judgment about a particular use of force, because that interval is shorter than the time a human needs to perceive, interpret, assess and authorize a lethal targeting decision. Conversely, where the temporal record shows that a human had a meaningful window to review and intervene, that window is itself a documented fact the tribunal can weigh against any contrary testimony.

[30] See Scientific Working Group on Digital Evidence, *Use of Probabilistic (AI) Outputs/Results in Digital Forensics*, AI Ad Hoc Committee draft, 15 January 2025, on file with the author, conditioning the use of machine-learning outputs on a validation showing (testability, known error rate, and reported accuracy) consistent with the reliability inquiry under Federal Rule of Evidence 702 and *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).

Throughput is the same point on a different scale. A system that produces 5,000 targeting recommendations per day will leave a temporal record that may be perfectly compliant on paper while making individualized human review functionally impossible. [31] The forensic question is not whether each recommendation was approved, but whether the rate at which approvals were issued is consistent with any plausible account of substantive human review.

4.3. The negative clarification

Evidence has limits. A forensic-ready system does not, on its own, attribute *mens rea*. The system is not a bearer of criminal guilt, however sophisticated, autonomous or fast. What it can do is preserve the artefacts from which the *mens rea* of human actors can be inferred, in the same way documentary evidence has always been used to prove what identifiable people knew, intended or recklessly disregarded. [32] The system is the evidentiary surface on which human knowledge and choice leave traces. The defendant remains human.

The concern commonly expressed in literature related to the responsibility gap for operators and commanders is directly relevant here. [33] If command responsibility becomes difficult because the machine is not a subordinate in the ordinary sense, the investigative focus has to shift to the human chain that authorized, configured, supervised and continued to rely on the system. Forensic records make that shift practical.

5. Digital Forensics as a Verification Methodology

Digital forensics, adapted from its long-established roles in domestic criminal evidence and corporate incident response, can serve as an inspection methodology for State compliance with rules governing AI-enabled force. Three conditions have to be met. First, the records that constitute the inspection target must be defined and standardized. Second, the methodology for collecting, authenticating and analyzing those records must be standardized. Third, the obligation to design systems so that they produce inspectable records must be embedded in procurement rather than left to post-incident discretion.

5.1. The standards that already do most of the work

A family of ISO/IEC standards, taken together, supplies a near-complete normative framework for the verification capability this paper proposes.

---

[31] L. Bainbridge, above note 26.

[32] Rome Statute, above note 6, Art. 30 (on the mental element).

[33] See Ann-Katrien Oimann, "Why Command Responsibility May (not) Be a Solution to Address Responsibility Gaps in LAWS", *Criminal Law and Philosophy*, Vol. 18, No. 3, 1 October 2024; Jonathan Kwik, "The Conceptual Roots of the Criminal Responsibility Gap in Autonomous Weapon Systems", *Melbourne Journal of International Law*, Vol. 24, No. 1, 2023.

ISO/IEC 27001 governs information-security management systems.[34] In the AI context, it supplies the institutional baseline under which evidence integrity, access to sensitive logs, and the auditability of authentication mechanisms are governed and audited; the operative technical controls sit in its Annex A and in ISO/IEC 27002.

ISO/IEC 27043 supplies the process model that holds the rest together.[35] It sets out the principles and processes of digital investigation, and it is the standard in which "readiness", being prepared to produce digital evidence before an incident occurs, is a defined pre-incident process class; it also locates chain of custody as a process running concurrently throughout an investigation, with governance of that readiness in turn the subject of ISO/IEC 30121. "Forensic readiness", as this paper uses the phrase, is that readiness process class applied to AI-enabled systems, not a coinage layered on top of the evidence-handling standards.

ISO/IEC 27037 governs the identification, collection, acquisition and preservation of digital evidence.[36] It is the operational-handling standard. Applied to AI systems, it requires that the artefacts identified in Section 4.1 be preserved using documented chains of custody, with authentication that can be tested by an independent examiner.

ISO/IEC 27042 governs analysis and interpretation of digital evidence.[37] It addresses the second half of the problem: once records exist and have been authenticated, how can their meaning be reconstructed in a way that is reproducible and contestable? Its discipline of separating fact from inference maps directly onto this paper's insistence that the system records conduct while the tribunal infers the mental element. In AI contexts, this includes the interpretation of model logs, confidence outputs and alignment-stage records, which often require domain expertise to read.

ISO/IEC 42001 is a management-system standard for artificial intelligence.[38] It does not itself prescribe a logging or evidence-production specification; it requires an organization to run a management system governing risk, transparency, monitoring, data provenance and impact assessment across the AI lifecycle, and it offers a reference catalog of controls, including the recording of event

---

[34] International Organization for Standardization, *ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection – Information security management systems – Requirements*, Geneva, 2022.
[35] International Organization for Standardization, *ISO/IEC 27043:2015 – Information technology – Security techniques – Incident investigation principles and processes*, Geneva, 2015 (defining the readiness process class and the concurrent chain-of-custody process); on governance of forensic readiness, see *ISO/IEC 30121:2015 – Governance of digital forensic risk framework*, Geneva, 2015.
[36] ISO/IEC 27037, above note 28.
[37] ISO/IEC 27042, above note 29.
[38] International Organization for Standardization, *ISO/IEC 42001:2023 – Information technology – Artificial intelligence – Management system*, Geneva, 2023.

logs, that the organization selects and declares through a statement of applicability.[39] It thus supplies the governance obligation and the control catalog from which a deploying State must derive and declare what it will record, while ISO/IEC 27037, 27042 and 27043 tell investigators how those records are handled, interpreted and fitted into an investigation. That 42001 leaves the specific record set to the organization's own selection is exactly why the procurement lever in Section 5.4 and the negative-disclosure duty in Section 9 are necessary.

These standards are voluntary unless adopted, but they are not parochial. They are the subject of established certification and assessment regimes in regulated domestic sectors such as finance, healthcare and critical infrastructure, and the institutional muscle to apply them already exists in many of the same States that field AI-enabled military systems.[40] What is missing is the explicit linkage between these standards and the international regimes governing armed conflict.

5.2. Three layers: preservation, authentication, interpretation

Reduced to its working parts, the verification methodology has three layers.

- *Preservation*: the system must generate and retain the records needed to reconstruct legally relevant events.
- *Authentication*: those records must be linked to reliable identities, times, model versions and chain-of-custody controls so that their integrity can be tested by an independent examiner.
- *Interpretation*: the records must be capable of being explained to legal decision-makers, with sufficient technical scaffolding that the proceeding does not collapse into an unreadable log dump.

A million log entries do not constitute verification if no one can connect them to the legal question. The layers draw on the standards discussed above, on the institutional baseline of ISO/IEC 27001 and within the process model of ISO/IEC 27043 throughout: preservation and authentication track the evidence-handling discipline of ISO/IEC 27037, and interpretation tracks the analysis discipline of ISO/IEC 27042, with ISO/IEC 42001 governing the AI management system that has to be designed to generate the records in the first place. Together they describe what a State has to do to make its AI-enabled operations legally inspectable at all.

---

[39] ISO/IEC 42001, above note 38, Annex A (notably control A.6.2.8, recording of event logs, and A.7.5, data provenance), the controls being selected and justified through the statement of applicability under Clause 6.1.3.

[40] For implementation evidence in regulated sectors, see European Union Agency for Cybersecurity (ENISA), *Cybersecurity Certification Statistics Report*, Heraklion, 2023.

5.3. Evidentiary Custodianship

The substantive claim a State seeks to make in any AI-related arms-control regime is some version of "we used force in a way that was within the rules". The verification claim that has to underwrite the substantive claim is more modest: "we can demonstrate, on the record, what we did and what we knew." This is what Evidentiary Custodianship names. A State exercises Evidentiary Custodianship with respect to a given rule when its systems and processes generate, retain and authenticate the records needed to prove compliance, and when it accepts that those records are not its private property but the evidentiary substrate of a regime to which it is bound, reviewable, under suitable confidentiality protections, by another State, an international body or a tribunal.

Evidentiary Custodianship is a procedural promise that the records exist and are reviewable. What it adds to ordinary forensic readiness is the second limb: not merely that a State keeps good records, but that it accepts in advance, as a regime commitment, that it no longer holds exclusive control over them. A State can be forensically ready and still treat its logs as sovereign property to be withheld; Evidentiary Custodianship is the posture that forecloses that move. It is what makes substantive promises testable. The posture is procedural in the strict sense: it does not assume that any particular operation was lawful; it assumes only that the question of lawfulness will be adjudicated against records the State has agreed in advance to produce. In this paper's argument, it is the strongest verification commitment that is technically achievable at present, and the floor below which AI-arms-control regimes cannot meaningfully operate.

5.4. Procurement as the practical lever

Most of this becomes operational at procurement.[41] A State that issues a procurement specification requiring AI-enabled targeting or decision-support systems to comply with ISO/IEC 27037, 27042 and 42001, including specific logging, retention and authentication requirements at training, alignment and runtime, has done more for accountability than a dozen high-level declarations. Procurement is the last point at which a State has unambiguous leverage over what a system is designed to do, and the first point at which subsequent operational secrecy and post-incident incentives have not yet hardened.

A State whose procurement specifications make no reference to forensic readiness has, by that omission, committed itself to a system whose accountability properties will be determined later, by other actors, under conditions less favorable to oversight. That is a choice with legal consequences. It cannot be undone by *post-*

[41] For the procurement-as-lever argument in a related context, see Yasmin Afina, *The Global Kaleidoscope of Military AI Governance: Decoding the 2024 Regional Consultations on Responsible AI in the Military Domain*, UNIDIR, Geneva, 2024.

hoc declarations. Anviksha Pachori's argument for "IHL-by-design" and continuous review converges with this point: design-stage requirements travel through the lifecycle far better than post-hoc commitments do.[42]

5.5. Auditable testing modules and pre-deployment evidence

The procurement obligation extends beyond logging in production. A forensic-ready system has to be testable before it is fielded. That is the second class of evidence the regime needs: auditable records of what the system was put through under controlled conditions, what failure modes were observed, and what those failures were classified as. Three categories deserve specific attention.

The first is adversarial testing and red-teaming. AI systems behave differently under deliberate stress, and the documented behavior under stress is itself evidence about the system's known limits. The NIST adversarial machine-learning taxonomy provides one vocabulary for classifying adversarial conditions, and the records of testing against that taxonomy are evidence about what was known.[43]

The second is protected-status testing. A system that can be made to misclassify a hospital, a religious site, a journalist or an *hors de combat* combatant has, in legal terms, a known specific-injury vector. The pre-deployment documentation of testing against protected-status edge cases, including the residual error rate after mitigation, is evidence about what the deploying State and its commanders accepted as a deployment-time risk. Dora Velenczei's work on auditable testing modules and on routing protected-status cues to human authorization is directly applicable. It converts a doctrinal worry about perfidy and improper exploitation of protected status into a documentation-and-routing problem that can be audited.[44]

The third is post-deployment validation. AI systems drift. The data they encounter in operations differs from the data they were trained on, and their behavior changes accordingly. A regime that requires only pre-deployment testing has a static view of a non-static object. Continuous monitoring records (observed inputs, classification distribution drift, confidence-output distribution drift, override frequency) are themselves part of the evidentiary spine. They are also closely related to the obligation of continuous review under Article 36, discussed below.

5.6. Verification is not (only) explainability

---

[42] Anviksha Pachori, "Governing AI-Integrated Weapons: Emerging Challenges of Accountability Under International Laws, *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.82-105.

[43] National Institute of Standards and Technology, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2 E2025, Gaithersburg, March 2025.

[44] Dora Velenczei, "The Perfidy Problem: Learned Deception in Lethal Autonomous Weapons Systems and the Limits of International Humanitarian Legal Compliance", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.1-28.

A common objection at this point is that AI explainability is contested, sometimes infeasible at the model level, and an inadequate basis for any legal regime. The objection has force, but it misses the proposal. Explainability asks whether a human can understand why a model produced a particular output. Verification asks whether a claim about what the system did, what it was designed to do, and what its operators knew can be tested against a record. A model may be difficult to explain internally while still producing verifiable external records: what data it received, what classification it produced, what confidence it displayed, what rule or threshold was applied, what human approval was sought, and what action followed. Conversely, an apparently explainable system may be unverifiable if its logs are incomplete or unauthenticated. The verification regime proposed here can and should draw on explainability techniques where they are mature, but it does not depend on full model interpretability to function.

6. Doctrinal Anchoring

The verification proposal does not displace any existing legal standard. It anchors them. The connection runs across four doctrinal pillars: meaningful human control and appropriate levels of human judgment, command responsibility, Article 36 weapons review, and Article 57 constant care.

A clarification on the source of obligation precedes the analysis. Additional Protocol I binds only its States Parties, and several States that field AI-enabled systems, including the United States whose "appropriate levels of human judgment" standard this paper engages, are not party to it.[45] The substantive principles invoked below, distinction, proportionality, and the constant care owed to the civilian population, bind all States as customary international humanitarian law regardless of ratification. The Articles of Additional Protocol I are cited throughout as the most precise codified expression of those obligations, not as their source for non-parties.

6.1. Meaningful human control and appropriate levels of human judgment

The phrase "meaningful human control" has been associated principally with civil-society advocacy and with the CCW process, where the International Committee of the Red Cross and others have given it shape.[46] Its core elements are that the operator must understand the system's functioning, must be able to predict and anticipate its effects, must retain the capacity to intervene or stop the system, and must preserve a

[45] On the customary status of distinction, proportionality and precautions in attack, binding irrespective of treaty ratification, see Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: Rules, ICRC and Cambridge University Press, Cambridge, 2005, Rules 1, 14 and 15. The United States is a signatory but not a party to Additional Protocol I. The Martens Clause independently subjects new means and methods of warfare to the principles of humanity and the dictates of public conscience: see *Legality of the Threat or Use of Nuclear Weapons*, above note 10, para. 87.
[46] ICRC, above note 2; Article 36 (NGO), above note 11.

sufficient connection between the human decision and the resulting use of force.[47] The phrase "appropriate levels of human judgment over the use of force" has been associated principally with United States Department of Defense Directive 3000.09, which treats the required level of human judgment as context-sensitive, calibrated to mission, target set, operational tempo and feasibility of intervention.[48]

The two standards do not need to be reconciled before they can be operated. They need to be made evidentiary. Each standard, in its own terms, generates concrete questions about the record. Did the operator have observability sufficient to understand the system's functioning? The interface logs and confidence-output records will speak to this. Could the operator predict and anticipate effects? The pre-deployment validation and known-failure documentation will speak to this. Was the connection between human decision and force sufficient, or was the human role calibrated to context appropriately? The control-plane configuration, approval-gate placement and temporal records will speak to this.

A forensic-ready system does not pick between meaningful human control and appropriate levels of human judgment. It supplies the records that allow either standard, applied by a tribunal or an inspector, to be tested rather than asserted.

6.2. Command responsibility and *mens rea*

Command responsibility, codified in Article 28 of the Rome Statute and domestically (for the Philippines) in section 10 of Republic Act No. 9851,[49] requires, in broad terms, proof that the accused had effective control over forces, that those forces committed the relevant crimes, and that the accused failed to take all necessary and reasonable measures within their power. Article 28 sets a different fault standard for each kind of superior: a military commander is liable where he "knew or, owing to the circumstances at the time, should have known" that the forces were committing or about to commit the crimes, while a civilian or other non-military superior is liable only where he "knew, or consciously disregarded information which clearly indicated" as much.[50] That distinction matters for AI-enabled force, because the human chain this paper traces runs through both kinds of superior. The Philippine vehicle is not identical: section 10 of Republic Act No. 9851 applies the single, stricter "knew or, owing to the circumstances at the time, should have known" standard to every superior, military and civilian alike, and does not adopt the Rome Statute's lower civilian threshold.[51] The case law on command responsibility (*Delalic et al.*

[47] ICRC, above note 2.
[48] US Department of Defense Directive 3000.09, above note 12.
[49] Republic Act No. 9851, above note 7, Sec. 10.
[50] Rome Statute, above note 6, Art. 28(a) (military commanders) and 28(b) (other superiors).
[51] Republic Act No. 9851, above note 7, Sec. 10, applies its fault standard to a superior with "effective command and control, or effective authority and control as the case may be," using the single formulation "either knew or, owing to the circumstances at the time, should have known," without the separate civilian-superior threshold drawn in Rome Statute Art. 28(b).

(*Celebici*), *Bemba*, *Blaskic* and others)[52] has worked through the elements in conventional-force settings. AI-enabled force does not displace the doctrine. It complicates how its elements are proved.

The question of effective control becomes, in part, a question about who configured the control plane, who set the mission parameters, who chose the oversight mode, and who retained override authority. Each of these is a person whose role can be reconstructed from authenticated records. The question of what the superior knew, should have known, or consciously disregarded becomes a question about what the training-stage documentation showed, what the alignment-stage records reinforced or relaxed, and what warnings were transmitted and received. These too are reconstructible from records, if those records were preserved.

The *mens rea* analysis is where the alignment stage carries the most weight. A State or a commander who reinforces, through deliberate fine-tuning, a model behavior that was already known to misclassify civilian objects has made a mental-state-relevant choice that no opacity claim can wash away. A State that procures a system that produces output volumes too high for individualized human review, having been warned of the resulting overtrust risk,[53] has done the same. As Kwik observes in the related context of sycophantic decision-support AI, perceived accuracy can be inflated by congruency, and military expertise does not automatically guard against this bias.[54] These are alignment-stage and training-stage facts. They are evidentiary, and they are reachable by forensic methods.

What forensic records do, in this setting, is allow a tribunal to distinguish among the mental states that command-responsibility doctrine treats as different. The case law from *Celebici* through *Bemba* turns on which fault standard the evidence proves. The operative categories are statutory: actual knowledge; the military commander's "should have known"; and the civilian superior's "consciously disregarded information which clearly indicated" (Article 28(a) and (b)

---

[52] International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Delalic et al. (Celebici)*, Case No. IT-96-21-A, Judgment (Appeals Chamber), 20 February 2001 (on knowledge and effective control); International Criminal Court, *The Prosecutor v. Jean-Pierre Bemba Gombo*, Case No. ICC-01/05-01/08, Judgment Pursuant to Article 74, 21 March 2016 (conviction subsequently reversed by Appeals Chamber, Judgment, 8 June 2018, which acquitted Bemba on the ground that the Trial Chamber erred in assessing the "necessary and reasonable measures" element, including the relevance of his position as a remote commander, and convicted beyond the scope of the confirmed charges; *Bemba* is cited here for its treatment of the elements of superior responsibility, which the acquittal did not displace, though the majority's treatment of the "reasonable measures" standard for remote commanders is itself debated); International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Tihomir Blaskic*, Case No. IT-95-14-A, Appeal Judgment, 29 July 2004.

[53] On the overtrust and automation-bias problem, see Kwik, below note 54, and the sources cited therein.

[54] Jonathan Kwik, "Digital Yes-Men: How to Deal with Sycophantic Military AI?", *Global Policy*, Vol. 16, No. 3, 2025, pp. 467–473, citing Anastasia Bashkirova and Dario Krpan, "Confirmation Bias in AI-Assisted Decision-Making: AI Triage Recommendations Congruent With Expert Judgments Increase Psychologist Trust and Recommendation Acceptance", *Computers in Human Behavior: Artificial Humans*, Vol. 2, No. 1, 2024, p. 100066.

respectively).[55] That *Bemba*'s conviction was later reversed on appeal, on the sufficiency of the measures he took rather than on the structure of the doctrine, does not weaken the point; if anything it underscores it, because the appellate disagreement was about what the evidence showed, which is precisely the question forensic records are meant to answer. Forensic categories map onto those fault standards with a usable degree of precision. A system fine-tuned after documented failure modes were flagged, with the safety thresholds relaxed in the same release, speaks to actual knowledge or to the conscious disregard of information that clearly indicated the risk. A workflow that rewards rapid approval of AI recommendations and produces approval rates inconsistent with substantive review speaks to what a commander, owing to the circumstances at the time, should have known. A user interface that suppresses uncertainty information that the underlying model produced speaks to the same standard, a design choice that made human control formally present but practically weak. A documented post-deployment drift that the deploying State could not reasonably have foreseen at procurement points away from fault altogether. None of these forensic patterns proves criminal responsibility on its own. They give courts, investigators and commanders a way to make the distinctions on which the doctrine actually depends.

The negative clarification needs reinforcement here. The system is the evidentiary surface, not the defendant. The *mens rea* attaches to people: commanders, operators, program managers, acquisition authorities, and in some circumstances developers, whose decisions about training, alignment, configuration, deployment and continued reliance on the system can be inferred from the records the system was designed to produce.

6.3. The duty to review under Article 36 of Additional Protocol I

Article 36 of Additional Protocol I requires States Parties, in 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 the Protocol or by other rules of international law.[56] The substantive prohibition the provision polices, that a State may not field a weapon whose normal use would violate international humanitarian law, reflects customary law; whether the

[55] The two standards are the fault elements specific to superior responsibility under Art. 28; they operate as exceptions to the default mental element of Art. 30 (intent and knowledge), which does not extend to recklessness or *dolus eventualis*. See Rome Statute, above note 6, Arts 28 and 30; International Criminal Court, *The Prosecutor v. Thomas Lubanga Dyilo*, Judgment Pursuant to Article 74, ICC-01/04-01/06, 14 March 2012, para. 1011 (declining to read *dolus eventualis* into Article 30, and not following the contrary view taken at the confirmation stage). Where this paper uses looser fault language drawn from domestic systems, it is by analogy to those systems, not as a freestanding category of international criminal law.

[56] AP I, above note 3, Art. 36.

procedural duty to conduct a formal review is itself customary is contested.[57] Its application to AI-enabled systems has at least one underappreciated dimension. A system whose operation cannot, after the fact, be reconstructed in evidentiary form is a system whose review is correspondingly weakened, because there is no methodology by which the State can verify, or external bodies can challenge, the conclusion that its operation is lawful. Forensic readiness is therefore what a genuine Article 36 determination of an AI-enabled weapon entails in practice, rather than an additional requirement read into the text. The argument does not rest on a novel reading of the provision. It rests on the observation that a State cannot meaningfully review what it cannot meaningfully reconstruct.

The reading also has a temporal dimension that conventional weapons did not require the law to take seriously. A conventional weapon, once reviewed, does not change. An AI-enabled system that learns, is fine-tuned in deployment, or is updated against new training data is not the same system after deployment as it was at the moment of review. A one-time review gate cannot discharge the underlying duty for such a system. The ICRC's own guidance on weapons review, and the continuous-review argument developed elsewhere in this special edition, treat review as something that must recur as the weapon materially changes;[58] honoring that approach requires documentation that tracks each material change to the model, the training corpus, the alignment configuration or the operational envelope. Forensic-ready systems are the precondition for review of that kind to be possible at all.

6.4. The constant-care obligation under Article 57

Article 57 of Additional Protocol I requires that "constant care" be taken to spare the civilian population in the conduct of military operations, and imposes specific precautionary obligations on those who plan or decide upon attacks.[59] Constant care is a continuing obligation. It does not exhaust itself at mission planning. Applied to AI-enabled systems, it requires that the operational use of the system be re-verified as the situation changes: target verification not left to the system alone, proportionality assessment revisited as new information arrives, and the lawful basis for an attack tested again before each engagement that the precautionary obligation reaches.

The constant-care obligation maps directly onto evidence categories. Re-verification produces records: the inputs at re-verification, the system's classification at that moment, the human's interface display, the confidence outputs, the time

---

[57] International Committee of the Red Cross, *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*, Geneva, January 2006. On the contested customary status of the procedural review duty as distinct from the substantive prohibition, see ibid. and the divergent State practice it surveys.

[58] ICRC, above note 57 (recommending review at successive stages of development and on modification); and Pachori, above note 42, on continuous review and IHL-by-design.

[59] AP I, above note 3, Art. 57.

available for decision. A regime that requires constant care without requiring the records that would prove constant care was exercised has, again, declared a duty without designing in the means to test it. Thanapat Chatinakrob's argument that war data should itself be treated as an object of protection sits comfortably alongside this conclusion. If data structures the conduct of hostilities, its preservation, protection and evidentiary treatment are not incidental.[60]

7. Case Studies

7.1. The 300-millisecond urban strike

Imagine that a targeting system classifies and acts within 300 milliseconds in an urban environment containing civilians. The temporal record, if preserved and if its timebase is itself authenticated, is decisive on the factual question of whether any real-time human review window existed. That qualification is not a formality: a sub-second interval is only as good as the timebase that produced it, a single monotonic clock of sufficient resolution stamping both events, or, where classification and approval are logged by different components, a recorded and bounded offset between their clocks. That is logging discipline that has to be designed in (the kind of event-logging control ISO/IEC 42001 leaves an organization to select), and whose acquisition-time counterpart ISO/IEC 27037 addresses when it requires the time source and any clock delta to be recorded. Subject to that, an interval this short is well below any documented human response time for perceiving, interpreting, assessing and authorizing a lethal action, so the record shows no window in which a human "on the loop" could have supplied real-time judgment about that engagement. Two things still belong to others. Whether the absence of a per-engagement window is decisive depends on where human authority was placed, since it may have been pre-delegated at activation rather than exercised at the moment of force; and whether the resulting architecture is sufficient under either standard is for the tribunal to decide. The forensic-readiness question is whether the temporal record, the operator-interface logs and the approval-gate settings exist, are authenticated and are producible. Where they do, they let the doctrinal questions be decided on the record rather than on assertion.

[60] Thanapat Chatinakrob, "War Data as an Object of Protection under International Humanitarian Law", in this special edition, Section 2.1, "Defining War Data and Its Functional Role"; Section 4.2, "Foreseeability, Feasibility, and the Epistemic Burden on Commanders"; and Section 5, "Accountability and the Future of Data Protection".

7.2. Phalanx and Iron Dome

Naval point-defense systems and missile-defense platforms operate in time envelopes that make per-engagement human approval unrealistic.[61] This is the case in which the appropriate-levels-of-human-judgment framing has its clearest application. The relevant human judgment may sit at activation, at rules-of-engagement configuration, and at mission-architecture design rather than at each intercept. The forensic-readiness contribution here is not to insist on an approval gate that could not exist. It is to show that even in this case verifiability is achievable. The control-plane settings, the activation records, the rules-of-engagement configuration and the per-intercept logs together permit a tribunal to reconstruct whether the operational use stayed within the bounded envelope that human authority defined. The case shows that forensic readiness does not require any one architecture of human involvement. It requires that whatever architecture was chosen leave a reviewable record.

7.3. Targeting pipelines at industrial scale

Large-scale AI targeting pipelines, of which Project Maven is a publicly discussed example,[62] present the verification problem at scale. A system that processes very large volumes of imagery and produces large numbers of targeting recommendations leaves humans nominally on the loop while making individualized review functionally impossible. The forensic question is not only whether per-recommendation approvals were issued, but whether the rate, distribution and dwell time of approvals are consistent with substantive human review at all. A State that procured such a pipeline without provisions for operator-attention metrics, throughput audit logs and pre-deployment overload modelling has, again, decided in advance what it is willing to be held accountable for. A State that procured it with such provisions has placed itself in a position to demonstrate, on the record, what the human role actually was.

7.4. Learned deception and protected cues

An AI-powered autonomous weapons system is trained partly in simulated environments. It learns, without any designer having explicitly programmed perfidy, that certain behaviors increase mission success by exploiting an adversary's restraint

---

[61] For technical descriptions of the point-defense and missile-defense systems referenced in this case study, see Congressional Research Service, *Defense Primer: US Ballistic Missile Defense*, IF10541, Washington, DC, updated 2024, available at: https://www.congress.gov/crs-product/IF10541 (all internet references were accessed on 30 May 2026); and the manufacturers' published descriptions of the Phalanx Close-In Weapon System (Raytheon, an RTX business) and Iron Dome (Rafael Advanced Defense Systems).

[62] Public reporting on Project Maven includes Kate Conger, "Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program", *Gizmodo*, 1 June 2018; on the program's later evolution, see public United States Department of Defense announcements concerning the Maven Smart System.

around protected cues.[63] For example, by approaching a structure that adversary forces are likely to treat as protected, then engaging when the adversary lowers its guard. During deployment, the behavior manifests, with civilian casualties.

The verification question is not only whether perfidy occurred, in the sense in which the law of armed conflict prohibits it. It is whether the risk was foreseeable, whether it was tested for, whether protected cues were routed to human authorization, and whether the State preserved an auditable record of what was done about the risk before the system was deployed. The relevant artefacts include simulation scenarios and reward-function configurations, red-team reports flagging exploit-of-restraint patterns, hard-block rules around protected cues, the alignment-stage records that determined whether such hard blocks were preserved or relaxed, and the runtime logs showing how protected cues were classified at the moment of the engagement. In a forensic-ready regime, each of these is evidence. In a regime without forensic readiness, the State's defense that the behavior was unforeseeable becomes difficult to test, and its inability to produce the records invites an adverse inference.

The case shows that forensic readiness is preventive as well as inculpatory. It is not only about proving guilt after harm. It is also about forcing institutions to identify, document and address foreseeable risks before deployment, and preserving the records that show whether they did.

8. Implementation Barriers

The argument so far has presented the verification proposal as if its only obstacle is conceptual. It is not. Six barriers will determine whether the proposal can become operational across the systems that already exist and the regimes that already bind States.

8.1. Mutual legal assistance is too slow

Cross-border evidence collection in conventional criminal matters relies on mutual legal assistance treaties (MLATs) and on letters rogatory, both of which operate on timescales of months to years.[64] AI-driven incidents unfold on timescales of seconds. The mismatch is structural. Verification regimes cannot rely on conventional MLAT routes for the kind of expedited, authenticated record exchange that AI-incident investigation will require. Some accelerated bilateral or multilateral protocol,

---

[63] See e.g. Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Lieber Institute West Point*, 5 December 2025; and Velenczei, above note 44, Part 3 (on the technical pathways by which AI-enabled systems learn to exploit protected-status cues).

[64] For the structural critique of mutual-legal-assistance timeliness in the digital age, see Andrew Keane Woods, "Against Data Exceptionalism", *Stanford Law Review*, Vol. 68, 2016, p. 729.

perhaps modelled on existing mutual recognition arrangements in cybersecurity incident response,[65] will need to be developed.

The shape of a workable solution is foreseeable. It would include pre-positioned evidence-preservation obligations triggered immediately on incident, defined channels for authenticated record exchange under controlled disclosure, and standing agreements about who may examine what records under what protections. The closest existing analogues are the 24/7 contact-point arrangements and the expedited preservation provisions of the Budapest Convention on Cybercrime. Neither was designed for AWS investigation, but both supply institutional muscle that could be adapted. The diplomatic problem is not easy. It is, however, genuinely a problem of protocol design on the back of an existing forensic-evidence regime, rather than a problem of inventing a forensic-evidence regime from scratch.

8.2. Judicial and prosecutorial education

A second barrier is interpretive. Records of model logs, alignment-stage decisions and confidence outputs are not self-explanatory to lawyers, prosecutors or judges who have not been trained to read them. Capacity-building inside legal institutions, analogous to what was required to make digital evidence routinely admissible in domestic criminal proceedings a generation ago, is a precondition for AI verification regimes to function. The Philippine Rules on Electronic Evidence,[66] adopted in 2001, anticipated some of this for a different evidentiary universe, and even there the framework reached criminal proceedings only by a later extension, which is itself a measure of how much institutional adjustment a new evidentiary category demands. The same kind of normative work has to be undertaken for AI-evidence categories internationally.

8.3. Storage, classification and data sovereignty

Forensic-ready AI systems generate large volumes of records. Secure storage is non-trivial. Classified handling may be required for some categories. Data-sovereignty rules may restrict cross-border transfer. These are legitimate operational and legal constraints. They do not invalidate the verification proposal, but they shape its design. Regimes will need to provide for in-country preservation with controlled foreign access, possibly via cryptographic commitments that can be opened only under defined conditions.[67] The technical primitives exist; the institutional architecture does not yet.

---

[65] See Council of Europe, *Convention on Cybercrime (Budapest Convention)*, ETS No. 185, 23 November 2001, Arts 16 (expedited preservation of stored computer data), 17 (expedited preservation and partial disclosure of traffic data) and 35 (24/7 network of contact points).

[66] Rules on Electronic Evidence, above note 18.

[67] For the cryptographic-commitment primitives applicable, see National Institute of Standards and Technology, *NIST FIPS 180-4: Secure Hash Standard*, August 2015; for institutional deployment patterns, see ISO/IEC 27001, above note 36.

8.4. The automated-forensics paradox

A particular tension recurs through the entire proposal. AI-scale logs may not be readable by humans operating at human speed. The interpretation of model behavior, especially of alignment-stage decisions and runtime classifications, may itself require AI assistance.[68] The verification of AI may depend on AI. The paradox is not fatal. It is, however, a reason to insist that the AI used in forensic analysis itself be subject to the same standards of transparency, validation and reviewability that the verification regime imposes on the systems it audits. A forensic AI that cannot be inspected fails the same test that the systems it inspects are required to pass.

8.5. Standards convergence

The standards relied on in this paper (ISO/IEC 27001, 27037, 27042, 27043 and 42001) have been developed through different working groups and on different timetables. Their effective use as a verification suite will require explicit cross-mapping, profile development for the AI-armed-conflict context, and recognition mechanisms across the States that adopt them. National AI risk-management frameworks, including those developed by NIST,[69] are partial analogues but were not designed with verification in mind.

A workable convergence would have three properties. First, an explicit AWS-domain profile of ISO/IEC 42001 that specifies, for AI systems used in armed-conflict contexts, the minimum logging, retention, authentication and validation requirements that a deploying State would have to meet. Second, a documented interface between that profile and the digital-evidence chain-of-custody requirements of ISO/IEC 27037 and 27042, so that the records produced under the AI management system can be handled and interpreted as digital evidence without losing authentication. Third, recognition mechanisms comparable to those used in cybersecurity certification schemes, under which States and international bodies could rely on each other's audit findings when assessing compliance, without requiring full re-audit at every stage.

8.6. Forensic readiness protects honest commanders too

A final point on acceptance. Forensic readiness is sometimes read as a one-sided imposition on militaries: more documentation, more audit, more scope for second-guessing. That reading is incomplete. Records cut both ways. A preserved log may show that the system remained inside a lawful engagement envelope; that the operator received and considered uncertainty warnings; that an override attempt was made; that the harmful classification was the product of an unauthorized post-

---

[68] This observation is the author's; on the adversarial conditions that make machine-generated behaviour difficult to interpret, see NIST, *Adversarial Machine Learning*, above note 43.
[69] NIST AI RMF, above note 21.

deployment modification rather than a baseline design choice; that a commander acted reasonably on the information available at the time. A forensic regime makes the legal inquiry more precise. It identifies what was foreseeable, what was controllable, what was known, what was preserved and what was done. That precision benefits lawful commanders as much as it benefits accountability. A regime that disciplines the inquiry into what actually happened is, in the long run, more protective of professional militaries than a regime in which every adverse outcome is litigated against gauzy declarations.

9. What a Verification Clause Should Require

The argument so far has been general. The minimum content of a verification clause, one that could be included in a future treaty on AI in armed conflict, in a unilateral or multilateral political declaration, in domestic procurement standards, or in military manuals and doctrine, comes to six elements.

First, lifecycle recordkeeping. States should preserve records from training, testing, alignment, procurement, deployment, updates and runtime operations, where those records are necessary to assess compliance with the relevant rule of international law or instrument.

Second, human-control traceability. States should preserve records showing where approval gates were located, who authorized mission parameters, what the operator saw, what warnings were displayed, whether override was practically possible, and whether human intervention occurred or was attempted.

Third, review continuity. A system materially changed through retraining, model update, sensor integration, mission expansion or threshold adjustment should trigger renewed legal and technical review. The review record should identify what changed and whether prior approval remains valid.

Fourth, tamper-evident preservation. Relevant records should be protected through access controls, hash verification, time-stamping, retention rules and chain-of-custody procedures consistent with ISO/IEC 27037 and the Philippine and other domestic electronic-evidence regimes.

Fifth, accountable non-preservation. Failure to preserve required records should carry legal or administrative consequences, especially where the absence of records prevents assessment of civilian harm, human control or command knowledge. Adverse inferences from non-preservation are not novel in evidence law. They operate in domestic criminal and civil proceedings on spoliation, and the same logic has a recognized inter-State form: where one party has exclusive control of the evidence, an international tribunal may resort more liberally to inferences of fact and

circumstantial evidence against it.[70] A deploying State has precisely that exclusive control over the records of its own AI-enabled operations.

A clarification on reach is owed here, because it bears on the mutual mistrust with which this paper opened. Forensic readiness and Evidentiary Custodianship are, in the first instance, things a State does to its own systems, a unilateral self-binding measure whose power to dissolve the mutual dilemma of Section 1 is contingent on reciprocal adoption, which the consensus history of the CCW process shows is the hard part. The adverse-inference logic above bites where a tribunal has jurisdiction, in a domestic prosecution under a statute such as Republic Act No. 9851, or between parties already bound by an instrument. The full mutual case therefore runs through the verification clause proposed below; what unilateral practice does in the meantime is establish the norms, standards and procurement patterns on which any later reciprocal regime would be built.

Sixth, negative disclosure. The procurement specification should require contractors to document, at the design stage, what the system does not log, retain or authenticate, and why. Without this requirement, a State can satisfy every positive logging duty while quietly omitting the categories that would, post-incident, be the most damaging to its position: operator-attention metrics, alignment-stage rollback decisions, override-attempt timestamps, prompt and interface history. After the fact, "we logged everything required" becomes a defense even when the unlogged categories are exactly the ones the legal inquiry needed. Negative disclosure converts that omission from a hidden choice into a procurement-stage commitment. The State has to declare, in advance and on the record, which categories it has decided to render unauditable. That declaration is itself evidence later, and, equally important, it gives reviewers a focal point for pushback before deployment, when leverage still exists. Negative disclosure is the structural counterpart to lifecycle recordkeeping, closing the gap that the positive obligation alone leaves open.

These requirements can and should be scaled. A small loitering munition, a ship-based defensive system, a strategic targeting platform and a foundation-model decision-support tool should not be forced into identical logging patterns. The relevant standard is functional. The greater the autonomy, lethality, operational complexity, civilian-risk profile and distance between human authorization and machine action, the stronger the forensic-readiness duty should be. A low-risk system may require modest records. A system capable of selecting or engaging targets in a civilian-dense environment should require a far richer evidentiary record. The same functional approach should apply to retention. Not every operational log must be preserved indefinitely. Records relevant to legal review, civilian harm, known failure modes, mission authorization, human approval and system action should have defined retention periods. Where civilian harm occurs, preservation should become

[70] International Court of Justice, *Corfu Channel Case (United Kingdom v. Albania)*, Merits, ICJ Reports 1949, p. 4, at p. 18 (a State's exclusive territorial control may permit the victim of a breach a more liberal recourse to inferences of fact and circumstantial evidence).

immediate and mandatory. Where a system is updated after an adverse incident, the prior version should not disappear, otherwise the update itself may erase the evidence needed to understand what went wrong.

The clause architecture also helps reconcile treaty necessity with treaty improbability. If a comprehensive AI-weapons treaty is politically unlikely in the short term, States, regional processes, journals and expert bodies can still develop forensic-readiness norms. Those norms can shape procurement, weapons review, military manuals, political declarations, domestic legislation and future treaty drafting. Huang and Yu's argument that the treaty-making process itself has normative value, even when a treaty is improbable,[71] is consistent with this approach. Forensic readiness gives the process a practical agenda: do not only debate what States should promise, but specify what evidence they must preserve so that the promise can be tested.

10. Conclusion

Forensic readiness is not engineering hygiene. It is the evidentiary architecture without which the legal framework governing AI-enabled force cannot be tested, and therefore cannot be relied on. Existing law (Article 36, Article 28, the precautionary obligations of Additional Protocol I, the Rome Statute, the domestic criminal-law provisions of States like the Philippines) already supplies the substantive constraints. What it does not yet supply is the verification capability that turns those constraints from declarations into evidence. Digital forensics, anchored in ISO/IEC 27001, 27037, 27042, 27043 and 42001, operationalized through procurement, and embedded in international cooperation arrangements, can supply that capability.

Three things follow if the argument is accepted. First, scholarship on AI-enabled force has to take verification seriously as a distinct problem. Discussion of meaningful human control, of appropriate levels of human judgment and of command responsibility cannot proceed as though the elements of these doctrines were self-evidencing. The records do not exist unless they are designed to exist. Second, States that wish to be taken seriously as restrained actors have to design their restraint to be inspectable. A declaration of compliance unbacked by reviewable records is, in the long run, an invitation to infer that the State did not regard the rule it claimed to follow as binding enough to leave a trace. Third, tribunals and inspection bodies have to insist on the records, and where records are absent, to draw the inferences appropriate to that absence. The Prisoner's Dilemma in armed-conflict AI cannot be solved by declarations of restraint: States that wish their restraint to be credible will have to make it inspectable, and States that wish to hold others to account will have to demand that the records exist. Anything else leaves the dilemma's pull intact.

---

[71] Huang and Yu, above note 8.

The proposal does not require new substantive obligations. It requires that existing obligations be designed in advance to be testable. That is what those promises mean now that software has entered the battlespace.

PART 5
REGIONAL PERSPECTIVES

AI in Defence: Implications for Southeast Asia Militaries

Erick Nielson C. Javier[*]

ABSTRACT

As the employment of military AI makes headlines in conflicts in Europe and the Middle East, Southeast Asian militaries too are showing increasing interest in the technology. This paper investigates the possible rationales and use cases for AI in select Southeast Asian militaries. AI in mission command, surveillance, and select weapons systems, may help meet territorial defence needs, as well as ensure technological relevance, and influence norms for military AI use at the regional level. However, technical, operational and materiel challenges will need to be addressed to fully realize AI's potential for Southeast Asian defence forces, while minimizing potential vulnerabilities.

***Keywords**:* military artificial intelligence, AI in defence, AI in warfare, military uses of AI, Southeast Asia military AI, AI for national defence

Introduction

Up until the mid-2010s, artificial intelligence (AI) was once considered to be the preserve of fiction. However, since the 2020s and with the explosion in popularity of learning language model (LLM) applications such as ChatGPT, Grok, Gemini, Anthropic and DeepSeek, private companies and governments alike are in a rush to incorporate AI in nearly every possible field of human endeavour. AI is everywhere, and warfare is no exception to this.

The application of AI to warfare has been the subject of intense academic and political debate, verging on cliche science-fiction tropes, as seen in many depictions on how AI will supposedly be used to kill all humans, or defend against threats that humans themselves could not prevail against. While the most fantastical visions of militarized AI, in the form of anthropomorphic and human-like "killer robot" characters or god-like intelligences leading enormous machine armies, may yet remain just fiction, there are many more practical uses for AI in warfare that are already being researched and implemented. Lethal autonomous weapon systems (LAWS) are perhaps the most-theorized application, eliciting the most concern from arms control and disarmament advocates. AI and automation more generally are also being investigated for important diagnostics, repair, maintenance, logistics and

[*]Erick Nielson C Javier is a Defence Research Officer in the Research and Special Studies Division of the National Defense College of the Philippines. His research interests include revolutions in military affairs, deterrence for small states and the future of warfare.

support functions.[1] The use of AIs to provide decision-making and targeting support have also become more controversial, following Israel's alleged use to generate target lists for its war in Gaza.[2]

While there are many developments and discussions on the use of AI in warfare, these have largely been within the context of great powers such as the United States of America, the Russian Federation, and the People's Republic of China. Despite some studies indicating interests in AI for defence among other Indo-Pacific countries, particularly those in Southeast Asia, there is a dearth of analyses on these states' underlying reasons and objectives. This paper investigates the possible rationales and use cases that would help explain why and how Southeast Asian militaries may try to leverage AI for defence.

For the purposes of this paper, "military AI" is defined as systems used "in operational military activities… directly relating to the planning, preparation and conduct of military operations." This includes AI applications in command and control; intelligence, surveillance and reconnaissance; weapons systems; and cyber and information systems."[3] Adopted from a February 2026 paper published by the Stockholm International Peace Research Institute (SIPRI), such a definition excludes other possible applications of AI in the defence establishment, including on human resource management, policy development and health services. This paper also focuses primarily on military AI use at the operational and tactical levels. This is currently where the vast majority of military AI research and deployment occurs; while there are potential uses for the AI at the strategic level,[4] this paper will not address them.

It should be acknowledged that what systems are considered AI vary considerably: ChatGPT and other popular LLMs are often labelled as generative AI, and are different from systems used specifically for weapons such as assisted aim sights, missile guidance, and fire control. There is some contention as to whether generative AI and other programs classified or marketed as "AI" can truly be

---

[1] *See* "AI-Driven Autonomy Revolutionizing Aircraft Inspections", *Lockheed Martin*, 28 March 2024, available at: https://www.lockheedmartin.com/en-us/news/features/2024/ AI-Driven-Autonomy-Revolutionizing-Aircraft-Inspections.html (all internet references were last accessed July 2026); Akhil Kadidal, "PLAAF deploys AI-assisted refuelling system for YY-20", *Jane's Defence Weekly*, 1 April 2026, available at https://www.janes.com/ defence-intelligence-insights/defence-news/defence/plaaf-deploys-ai-assisted-aerial-refuelling-system-for-yy-20-tanker.

[2] Yasmeen Sehran, "How Israel Uses AI in Gaza—And What It Might Mean for the Future of Warfare", *TIME*, 19 December 2024, available at: https://time.com/7202584/gaza-ukraine-ai-warfare/.

[3] Netta Goussac and Vincent Boulanin, *Responsible Procurement of Military Artificial Intelligence*, Stockholm International Peace Research Institute, February 2026, p. 2, available at: https://www.sipri.org/publications/2026/other-publications/responsible-procurement-military-artificial-intelligence.

[4] 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, 2025, p. 11, available at: https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf.

classified as "intelligence", due to their limited ability to understand and infer abstract concepts.[5] For the purposes of this paper's analysis, they will be considered as "AI". As far as is known, all extant applications of AI are not artificial general intelligence (AGI), or strong AIs that can actually replicate human intelligence.[6]

What is and why AI in defence?

Applying direct force

AI in defence is generally considered an enabling technology, rather than a standalone system.[7] This means that there are many potential real-world applications for AI in defence. In terms of weapon systems specifically intended with AI in mind, LAWS has animated years of high-level discussions under the ambit of the Convention on Certain Conventional Weapons (CCW). In the December 2025 version of the rolling texts of the United Nations Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE-LAWS), LAWS are described as "functionally integrated combinations of one or more weapons and technological components, that can identify, select and engage a target, without intervention by a human operator in the execution of these tasks."[8] Prior to this, the United States' Department of Defense defined autonomous weapons systems as a "weapon system that, once activated, can select and engage targets without further human intervention by an operator."[9] The common denominator in these definitions or descriptions is the ability to select and engage targets without intervention by a human. Such systems promise removing ones' own soldiers from risk, and provide reaction times faster than that possible with manual inputs, which may be critical in certain combat scenarios that require above-human reflexes, such as air and missile defence.[10] However, the prospect of giving weapons the ability to decide to take lives, perhaps even without humans supervising or giving orders, is a major operational

---

[5] John Nosta, "Can LLMs Think Like Us?", *Psychology Today*, 24 August 2024, available at: https://www.psychologytoday.com/us/blog/the-digital-self/202408/can-llms-think-like-us.

[6] B. Jack Copeland, "Artificial Intelligence", *Britannica*, 02 February 2026, available at: https://www.britannica.com/technology/artificial-intelligence.

[7] N. Goussac and V. Boulanin, above note 3.

[8] United Nations Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Box I, Item 1, 18 December 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2026)/CCW_GGE_LAWS_Rolling_Text_-_status_18_December_2025.pdf.

[9] United States Department of Defense Directive 3000.09 – Autonomy in Weapon Systems, 25 January 2023, p. 22, available at https://media.defense.gov/2023/Jan/25/ 2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF.

[10] John K. Hawkley, *Patriot Wars: Automation and the Patriot Air and Missile Defense System*, Voices from the Field Research Report, Center for a New American Security, January 2017, p. 4, available at: https://www.jstor.org/stable/resrep06103?seq=8.

and ethical issue, and why LAWS garners much concern from futurists, philosophers and legal luminaries.[11]

Weapon systems do not need to be designed bespoke to be able to make use of AI, as they are now being added as an upgrade or improvement to existing weapons categories without necessarily replacing human operators. The manned F-35 fighter is an example of an existing platform being upgraded with AI; under Project Overwatch, onboard AI will assist pilots in identifying contacts detected by the aircraft's sensors.[12] There are also so-called "smart sights", or assisted optical aiming systems, like the SMASH Handheld developed by the Smart Shooter company, which uses "artificial intelligence, computer vision and machine learning"[13] in a module that can be attached to any rifle or small arm to improve aim regardless of conditions, greatly improving the effectiveness of soldiers using the system. Another example of a "platform-agnostic" AI, that can be installed on different platforms, is the Hivemind "AI pilot" developed by the Shield AI company, which has flown a variety of aircraft types, from quadcopter drones, larger unmanned air vehicles, to even the F-16, a manned fighter aircraft.[14]

Nor is AI's contribution to defence and warfighting limited to being part of weapons. AI and automation could be conceivably applied to any of the following major warfighting functions: mission command, movement and manoeuvre, intelligence gathering and processing, fires, sustainment, and protection.[15] Each function encompasses multiple tasks, not all of which require direct application of force. Such activities that do not directly involve target selection, target engagement and direct application of military force, are also known as "upstream" military tasks.[16] There are different desirable levels of autonomy corresponding to the specific taskings required, with particular emphasis on whether it is computationally

---

[11] William Boothby, "Dehumanization: Is There a Legal Problem Under Article 36?", in Wolff Heinschel von Heinegg, Robert Frau and Tassilo Singer (eds), *Dehumanization of Warfare – Legal Implications of New Weapons Technologies*, Springer International Publishing AG, Cham, 2018, pp. 21-51; Netta Goussac, "Safety net or tangled web: Legal reviews of AI in weapons and war-fighting", *Humanitarian Law and Policy*, 18 April 2018, available at: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/.

[12] Sydney J. Freedberg, "Lockheed test-flies F-35 with artificial intelligence to quickly ID unknown contacts", *Breaking Defense*, 26 February 2026, available at: https://breakingdefense.com/2026/02/lockheed-test-flies-f-35-with-artificial-intelligence-to-quickly-id-unknown-contacts/.

[13] "About", *Smart Shooter*, 2026, available at: https://www.smart-shooter.com/about-smart-shooter/.

[14] Colin Demarest, "Shield AI to let Hivemind software fly three more aircraft", *DefenseNews*, 9 April 2024, available at: https://www.defensenews.com/artificial-intelligence/2024/04/08/shield-ai-to-let-hivemind-software-fly-three-more-aircraft/.

[15] Thomas Ryan and Vikram Mittal, "Potential for Army Integration of Autonomous Systems by Warfighting Function", *Military Review*, Vol. 99, No. 5, September-October 2019, p. 123.

[16] Sarah Grand-Clément, *Artificial Intelligence Beyond Weapons - Application and Impact of AI in the Military Domain*, UN Institute for Disarmament Research, Geneva, 2023, p. 10, available at: https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf.

expansive and taxing on mental endurance.[17] The intelligence, sustainment and protection warfighting functions were deemed to be the largest beneficiaries of increased application of autonomous systems[18], as these functions normally need large numbers of personnel and high responsiveness which could be more seamlessly provided via autonomous systems.

The impact of using AI for intelligence can be seen in the Israeli use of AI as decision support systems (DSS) for its operation during the war in Gaza.[19] Examples include *Habsora* (Gospel), which was allegedly employed from 2023-onward,[20] and Lavender, which began to appear in reports from March-April 2024.[21] These systems aggregate and analyse data from multiple information streams to identify and recommend potential targets to human analysts, who then pass the information directly to combat commanders in the field. *Habsora*, Lavender, and other companion AI have allowed the rapid generation of tens of thousands of targets for Israeli strikes during the Gaza war, up to fifty (50) times faster[22] than could be possible with teams of human intelligence analysts.

The data used to generate these targets, while not publicly confirmed, are generally assumed to include image intelligence (IMINT) data from satellites and drones, and electronic and radio data from other sources such as smartphone messages, phone calls and video feeds.[23] The sheer volume and variety of data gathered is so expansive that it could actually hobble a military's ability to gain actionable insights from such data, making AI necessary to parse through the information.[24] While Israeli AIs enable the rapid identification of valid military targets such as missile launchers and arms, the large-scale civilian casualties from the attacks, which were supposedly recommended by Israeli AI DSS, have triggered questions on the exact parameters that Israeli AI use to classify targets. As AI depend

[17] T. Ryan and V. Mittal, above note 15, p. 124.
[18] T. Ryan and V. Mittal, above note 15, p. 133.
[19] Marta Bo and Jessica Dorsey, "Symposium on Military AI and the Law of Armed Conflict: The 'Need' for Speed – The Cost of Unregulated AI Decision-Support Systems to Civilians", *OpinioJuris*, 4 April 2024, available at: https://opiniojuris.org/2024/04/04/ symposium-on-military-ai-and-the-law-of-armed-conflict-the-need-for-speed-the-cost-of-unregulated-ai-decision-support-systems-to-civilians/.
[20] Duval Abraham, "'A mass assassination factory': Inside Israel's calculated bombing of Gaza", *+972 Magazine*, 30 November 2025, available at: https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/.
[21] Vinay Patel, "Israeli AI Technology Used to Identify 37,000 Hamas Targets", *International Business Times*, 4 April 2024, available at: https://www.ibtimes.co.uk/israeli-ai-technology-used-identify-37000-hamas-targets-1724204.
[22] D. Abraham, above note 20.
[23] Anwar Mhajne, "Regulating Military AI: Gaza and Beyond", *Survival*, Vol. 68, No. 2, May 2026, p. 26.
[24] Noah Sylvia, *The Israel Defense Forces' Use of AI in Gaza: A Case of Misplaced Purpose*, Royal United Services Institute, London, 04 July 2024, available at: https:// www.rusi.org/explore-our-research/publications/commentary/israel-defense-forces-use-ai-gaza-case-misplaced-purpose.

heavily on the quality of training data used to identify patterns, fears have surfaced[25] that the targeting metrics may be based on biased variables that have no direct bearing on whether the target in question is a legitimate member of the adversary organization, let alone combatants.

Speed in decision-making

The desire for fast reaction across the various warfighting functions follows from preexisting trends for more automated weapon systems that progressively reduce the numbers of humans needed for functions such as loading and maintenance. A similar rationale inspires the implementation of AI in upstream military tasks outside of weapons systems, especially in intelligence gathering and processing algorithms that improve the speed of threat detection, prioritization and decision-making for command and control. This is part of a long-standing desire to enhance a military's observe-orient-decide and act (OODA) loop. First coined by US Air Force Colonel John Boyd, it is a type of action-learning cycle[26] that asserts that the faster one can observe, orient, decide then act compared to their adversaries – "getting inside the enemy's OODA loop" – the better the chances of winning. Superior decision-making speed will help a military react faster than their adversaries, allowing them to gain or retake initiative and maintain pressure until the enemy capitulates and is defeated.

Great powers lead the way

This "need for speed" is clearly seen in the United States' Artificial Intelligence Strategy, published by the Department of Defense (or Department of War) on January 2026, which stated that "[w]e must weaponize learning speed, and measure and manage cycle time and adoption rates as decisive variables in the AI era."[27] This builds on earlier US policies on incorporating AI into defence as part of the Third Offset Strategy[28], which highlighted how emerging technologies can be used to offset the growing conventional and asymmetric capabilities of both state competitors and non-state adversaries. To this end, the United States works extensively to implement AI tools across the armed forces, from dedicated platforms such as the Collaborative

[25] Geoff Brumfiel, "Israel is using an AI system to find targets in Gaza. Experts say it's just the start", *National Public Radio*, 14 December 2023, available at: https://web.archive.org/web/20240220015604/https://www.npr.org/2023/12/14/1218643254/israel-is-using-an-ai-system-to-find-targets-in-gaza-experts-say-its-just-the-st.

[26] Brett Crowley, "The OODA Loop", *The Decision Lab*, 2025, available at: https://thedecisionlab.com/reference-guide/computer-science/the-ooda-loop.

[27] United States Department of War Artificial Intelligence Strategy, 9 January 2026, p. 4, available at: https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.

[28] United States Department of Defense, "The Third U.S. Offset Strategy and its Implications for Partners and Allies", 28 January 2015, available at: https://www.defense.gov/News/Speeches/Speech/Article/606641/the-third-us-offset-strategy-and-its-implications-for-partners-and-allies/.

Combat Aircraft (CCA) programs of the US Air Force[29], to backfitted solutions to improve the readiness and capability of existing military forces such as Project Overwatch and SMASH Handheld mentioned earlier. The 2026 AI Strategy called for further integration of AI systems in the US military's upstream tasks, exemplified in the development and deployment of the Maven Smart System. Developed under Project Maven, it is an expansive data analysis and decision-making software that is intended to be able to fuse data from a variety of sources and interoperate with other US systems.[30] The Maven Smart System has already seen wide use in strike planning, notably during the US-Israeli air campaign against Iran, which began on 28 February 2026.[31]

For the People's Republic of China and its People's Liberation Army, their drive to AI in defence is seen in its concepts of fighting "informatized local wars", wherein information is considered a key domain of warfare and one of, if not the, central means to wage conflict[32] in the modern era. Indeed, PLA strategic writings emphasize information sources, quality of collection and circulation speed among other variables to "seek decision-making advantages with information advantages".[33] China recognizes the importance of developing AI among other cutting-edge technologies for military advantage, as stated in its 2019 Defence White Paper.[34] Incorporating these cutting-edge technologies is key to PLA modernization, representing the third of a three-phase program of "mechanization, informatization and intelligentization".[35] To meet these targets, Chinese developers have demonstrated an array of AI-enabled technologies, from Norinco's Intelligent Precision Strike System that uses AI to deploy and control multiple drones to find

---

[29] Jennifer DiMascio, *US Air Force Collaborative Combat Aircraft*, Congressional Research Service, 28 November 2025, p. 1, para. 1, available at: https://www.congress.gov/crs-product/IF12740.

[30] Courtney Albon, "Palantir wins contract to expand access to Project Maven AI tools", *DefenseNews*, 31 May 2024, available at: https://www.defensenews.com/artificial-intelligence/2024/05/30/palantir-wins-contract-to-expand-access-to-project-maven-ai-tools/.

[31] Sydney J. Freedman, "'Insatiable appetite' for AI: Maven usage surged for strikes on Iran, Pentagon AI chief says", *Breaking Defense*, 12 May 2026, available at: https://breakingdefense.com/2026/05/insatiable-appetite-for-ai-maven-usage-surged-for-strikes-on-iran-pentagon-ai-chief-says/.

[32] Edmund J. Burke *et al.*, *People's Liberation Army Operational Concepts*, RAND Corporation, 2020, p. 5, available at: https://www.rand.org/content/dam/rand/pubs/ research_reports/RRA300/RRA394-1/RAND_RRA394-1.pdf.

[33] Xiao Tianliang *et al.*, "Chapter 13 - Operational Guidance", *Science of Military Strategy*, National Defense University Press, Beijing 2020, p. 227. Translation available at: https://www.airuniversity.af.edu/CASI/Display/Article/2913216/in-their-own-words-2020-science-of-military-strategy/.

[34] State Council of the People's Republic of China White Paper on China's National Defense in the New Era, 24 July 2019, pp. 6-7, available at: https://english.www.gov.cn/atts/stream/files/5d3943eec6d0a15c923d2036.

[35] Josh Baughman, "The Path to China's Intelligentized Warfare: Converging on the Metaverse Battlefield", *Cyber Defense Review*, US Army Cyber Institute, West Point, 16 December 2024, p. 30, available at: https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/4012231/the-path-to-chinas-intelligentized-warfare-converging-on-the-metaverse-battlefi/.

targets and devise and execute attacks, AI-enabled synthetic brigades that combine AI with multiple manned and unmanned platforms, and proprietary operational decision support, intelligence repository and battlefield intelligence systems, meant to showcase China's growing capabilities in the military AI domain.[36]

As both the United States and China are considered the leaders in military AI and its applications, and are currently locked in an intense geopolitical competition, it is very likely that this drive for speeds and wide applicability will inform other militaries' efforts to adopt AI.

ASEAN and AI in defence

States' efforts to incorporate military AI

Figure 1. ASEAN 5 Military AI Applications by Warfighting Area as of March 2026
Source: Author's illustration

The implications for AI in defence have not been lost in the Southeast Asian region. Individual Southeast Asian states have already begun investigating the potential uses of AI in defence for several years prior. Singapore has been leading the way among its peers, in keeping with its established practices of maintaining strong defence

[36] Tye Graham and Peter W. Singer, "New Products Show China's Quest to Automate Battle", *Defense One*, 2 March 2025, available at: https://www.defenseone.com/threats/ 2025/03/new-products-show-chinas-quest-automate-battle/403387/.

capabilities.[37] The Singapore Armed Forces (SAF) was among the first ASEAN member militaries to begin testing applications of AI in mission command, using a command and control information system (CCIS) fitted with AI in the Forging Sabre series of military exercises with the United States since 2021.[38] In 2022, it formed the Digital and Intelligence Service (DIS), effectively a fourth branch of the SAF, whose tasks include optimizing the application of AI across the entire spectrum of the SAF's operations.[39] In early 2026, SAF began testing unmanned aerial vehicles (UAV) integrated with Shield AI's Hivemind SDK AI system, ostensibly for intelligence and surveillance missions.[40]

The Indonesian Ministry of Defence announced in 2024 that it would cautiously explore adopting AI systems, particularly for "securing airspace", although they acknowledged that the technology was far from reliable[41]; the Indonesian Air Force in particular pushing for the technology's adoption. The Indonesian Navy have begun testing KSOT-008 torpedo-armed AI-powered autonomous underwater vehicles in October 2025,[42] as part of Indonesia's drive for self-reliance and their ongoing naval modernization.

The Malaysian Ministry of Defence, in its mid-term review of its 2019 Defence White Paper, highlights the need to acquire the latest technologies, such as AI tools.[43] The Malaysian Armed Forces have apparently already begun doing so, stating in parliamentary inquiries that they have been integrating AI tools for surveillance in land, air and sea domains, as well as using AI-assisted training

---

[37] Daniel Lu, *Poison Shrimp: Deterrence, Defense Cooperation, and Total Defense from Europe and Asia*, Naval Postgraduate School, December 2020, p. 29, available at: https://calhoun.nps.edu/server/api/core/bitstreams/c13d876a-7df8-4cbb-a919-c25f7461fa1 1/content.

[38] Aqil Hamzah, "AI, data analytics improve decision-making process, but human touch still key at SAF exercise", *Straits Times*, 22 September 2023, available at: https:// www.straitstimes.com/singapore/ai-data-analytics-improve-decision-making-process-but-human-touch-still-key-at-saf-exercise.

[39] Ministry of Defence of the Republic of Singapore, "Remarks By Minister For Defence Dr Ng Eng Hen at the Aspen Strategy Group and Amazon Web Services Roundtable On 'Allied Artificial Intelligence: Strengthening The Technological Edge", 14 February 2025, available at: https://www.mindef.gov.sg/news-and-events/latest-releases/15feb25_speech 2/.

[40] Oishee Majumdar, "Singapore tests UAV equipped with Shield AI software", *Janes Defence Weekly*, 06 March 2026.

[41] Walda Marison and Resinta Sulistiyandari, "Indonesia to cautiously adopt AI in defense by 2025", *Antara Indonesian News Agency*, 19 December 2024, available at: https:// en.antaranews.com/news/338622/indonesia-to-cautiously-adopt-ai-in-defense-by-2025.

[42] Akhsan Erido Elezhar, "Indonesia Debuts AI-Powered Autonomous Submarine Armed With Six Torpedoes", *The Defense Post*, 10 October 2025, available at: https:// thedefensepost.com/2025/10/10/indonesia-ai-submarine-torpedoes/.

[43] Adam Leong Kok Wey, "Malaysia's Defense White Paper Midterm Review: Emerging AI-Enabled Security Threats", *The Diplomat*, 23 September 2025, available at: https:// thediplomat.com/2025/09/malaysias-defense-white-paper-midterm-review-emerging-ai-enabled-security-threats/.

simulators and logistics systems.[44] There also seem to be intentions to implement "AI-enabled military drones, enhanced cyber defence and smart weapon automation."[45]

The Philippines, in its National Security Policy 2023-2028, stated that they would "innovate and modernize maritime, air, and space systems and security operations through the use of space science and technology applications, and integration of AI."[46] The Philippine Department of National Defense acknowledged the threat posed by AI systems used by their adversaries,[47] which was evidenced in the release of an internal memorandum banning the use of AI image generator apps which were seen as potential vectors for data breaches and data gathering for deepfake operations.[48] The Armed Forces of the Philippines (AFP) are studying the potential uses of AI in defence, including trial applications of systems like ChatGPT,[49] and have procured some AI-powered systems such as the SMASH Handheld rifle sight for counter-unmanned aerial system (C-UAS) work, which debuted publicly[50] during the 2025 *Balikatan* Exercises between the AFP and the United States Armed Forces.

The Royal Thai Air Force (RTAF) in its 2024 Defence White Paper outlined plans to develop AI – specifically an RTAF equivalent of ChatGPT, beginning in 2028.[51] Such an initiative is intended to help the RTAF in its ambitions to be a future-ready air force in the region, and is in line with other Thai military research that investigated the potential uses of generative AI for defence tasks.[52] The 2024 White Paper also indicated RTAF initiatives to develop unmanned combat air vehicles (UCAV), tactical UAVs, and manned-unmanned teaming (MUM-T) as part of the force's future expansion and modernization.[53]

---

[44] Nurnabihah Dini, "Malaysia advances smart military transition with new cyber, AI capabilities", The *Malaysian Reserve*, 19 November 2025, available at: https:// themalaysianreserve.com/2025/11/19/malaysia-advances-smart-military-transition-with-new-cyber-ai-capabilities/.
[45] *Id*.
[46] National Security Policy 2023-2028, National Security Council of the Philippines, p. 20.
[47] JV Ordonez, "Why The DND Wants a Revamped Military Modernization Plan", *Esquire Magazine*, 31 July 2025, available at: https://www.esquiremag.ph/politics/news/dnd-seeks-revamped-modernization-plan-to-counter-cyberthreats-a8072-20250731.
[48] Erick Nielson C Javier, "Artificial intelligence threat makes data protection priority for Philippine military", *East Asia Forum*, 3 July 2024, available at: https://eastasiaforum.org/ 2024/07/03/artificial-intelligence-threat-makes-data-protection-priority-for-philippine-military/.
[49] Jekki Pascual, "AFP studying use of AI for defense", *ABS-CBN News*, 17 January 2025, available at: https://www.abs-cbn.com/news/technology/2025/1/17/afp-studying-use-of-ai-for-defense-1553.
[50] Pia Lee-Brago, "Modern weapon systems showcased at Balikatan drills", *Philstar Global*, 05 May 2025, available at: https://www.philstar.com/headlines/2025/05/05/2440569/ modern-weapon-systems-showcased-balikatan-drills.
[51] Royal Thai Air Force (RTAF) Defence White Paper 2024, pp. 31 and 74.
[52] Kanok Bunnag, "Generative Artificial Intelligence and its Military Application", *Defence Technology Academic Journal*, Vol. 6, No. 13, pp. 28–41, May 2024.
[53] RTAF, above note 51, pp. 46-48.

The above is not an exhaustive review of these Southeast Asian states' official statements or forays into military AI applications. Nor is it meant to be a review of all ASEAN military AI capabilities, as efforts from Brunei, Cambodia, Laos, Myanmar, Timor Leste and Vietnam have yet to be reliably documented in the open media. It should be sufficient to illustrate the real interest and efforts of states in the region to harness the technology for their defence and security.

ASEAN tackles AI

In addition to individual member states, AI in defence has increasingly gained the attention of ASEAN as a collective. Interest in the topic could be evidenced as early as 2021, with Track II[54] discussions on the impacts of AI and unmanned systems to defence and security for the Southeast Asian region.[55] This was followed by a Track II workshop held in 2025 focused on AI in defence. In February 2025, the ASEAN Defence Ministers Meeting (ADMM) released the Joint Statement by the ASEAN Defence Ministers' Meeting on Cooperation in the Field of the Defence Sector,[56] marking the first regional statement specifically on the issue of AI in defence.

Prior to ADMM and Track II discussions on ASEAN, the organization released the ASEAN Guide to AI Governance and Ethics in 2024. The guide skirted the issue of military AI, as it was intended to be a "practical guide for organizations in the region that wish to design, develop and deploy traditional AI technologies in commercial and non-military or dual-use applications."[57] An ASEAN Working Group on AI Governance was thereafter also established.

[54] Founded in 2007, the Network of ASEAN Defence and Security Institutions (NADI) is a Track II network of ASEAN defence academic institutions that also serve as think-tanks. While not encompassing the totality of ASEAN's defence intelligentsia and think-tank population, its status as a recognized Track II platform that directly forwards information and recommendations to the ASEAN Defence Ministers' Meeting (ADMM) gives it some weight in representing views of these states' defence establishments with regard to thinking about AI in defence, even if some of the ASEAN states have not yet publicly disclosed specific defence AI programmes or experiments.

[55] Chairman's Report of the Track II NADI Workshop on Emerging Technologies and Its Impacts to Defence and Security, 29-30 September 2021, available at: https://www.rsis.edu.sg/nadi/pdfs/emgtech_1/Chairman_Report.pdf.

[56] Joint Statement of the ASEAN Defence Ministers on Cooperation in the Field of Artificial Intelligence in the Defence Sector, 26 February 2025, available at: https://asean.org/wp-content/uploads/2025/02/JS-ON-COOPERATION-IN-THE-FIELD-OF-AI-IN-THE-DEFENCE-SECTOR.pdf.

[57] ASEAN Guide on AI Governance and Ethics, February 2024, p. 3, available at: https://asean.org/wp-content/uploads/2024/02/ASEAN-Guide-on-AI-Governance-and-Ethics_beautified_201223_v2.pdf.

Analysis of ASEAN's interest in AI

Following trends and meeting threats

Southeast Asian views on AI in defence seem to largely follow global discussions on the technology. They generally acknowledge AI's potential to increase speed and efficiency in defence and security, as well as the ethical risks. There is some perception among ASEAN member states that AI will be useful for not just warfighting, but for non-combat military tasks such as humanitarian assistance and disaster response (HADR), as well as general surveillance. At least one Southeast Asian state sees AI as useful in supplementing military anti-corruption practices,[58] which while important, do not neatly fall in any of a military's warfighting functions. For another ASEAN member state, the importance placed on AI was not matched with substantive discussion on what its warfighting applications would be.[59]

Examining some of the stated rationales from ASEAN government officials, a case could also be made that the interest of Southeast Asian states to incorporate AI in defence is as much about ensuring they are not left out in the discussion and exploitation of this technology. Statements in both Track I[60] and Track II[61] defence discussions within ASEAN generally echo the perception of AI being an ubiquitous technology that is being harnessed everywhere, therefore individual member states and indeed ASEAN as a whole must have access to the technology too. There is general sentiment among ASEAN members that "halting AI is not an option, and instead, ASEAN must learn to adapt while addressing disparities in AI capacities across the region."[62]

In addition to access to the technology, some ASEAN states have exerted active efforts to influence the shaping of international regulations on military AI. Lao PDR, the Philippines and Singapore participate in the GGE-LAWS discussions,

---

[58] Bernama, "Mindef to leverage AI, digital tech to tackle corruption in armed forces", *New Straits Times*, 31 January 2026, available at: https://www.nst.com.my/news/nation/2026/01/ 1368695/mindef-leverage-ai-digital-tech-tackle-corruption-armed-forces.

[59] Kuik Cheng-Chwee, *Malaysia Reassesses Security: The Mid-Term Review of the Malaysian Defence White Paper*, Perspective Issue No. 11-2026, ISEAS-Yusof Ishak Institute, Singapore, p. 4, available at: https://www.iseas.edu.sg/wp-content/uploads/ 2026/01/ISEAS-Perspective_2026_11.pdf.

[60] Joint Statement of the ADMM on Cooperation in the Field of Artificial Intelligence in the Defence Sector, p. 2, available at: https://asean.org/wp-content/uploads/2025/02/JS-ON-COOPERATION-IN-THE-FIELD-OF-AI-IN-THE-DEFENCE-SECTOR.pdf.

[61] Chairman's Report of the Track II NADI Meeting on The Emerging Security Architecture in Asia-Pacific, Assessment of Implications, and Recommendations for ASEAN, 18-21 September 2024, p. 1, available at: https://www.rsis.edu.sg/nadi/pdfs/ESAAPAIR_1/ ChairmanReport.pdf.

[62] Chairman's Report of the Track II NADI Meeting on Artificial Intelligence (AI) in Defence: Potential and Challenges for ADMM Cooperation, 26-29 August 2025, p. 10, available at https://www.rsis.edu.sg/nadi/pdfs/AIiD_1/ChairmanReport.pdf.

while Thailand is an observer.[63] The Philippines co-endorsed a draft protocol VI for the CCW in the 2022 sessions of GGE-LAWS,[64] and also organized the Manila Meeting on Indo-Pacific Perspectives on Autonomous Weapon Systems on 13-14 December 2023.[65] Singapore co-hosted the Responsible AI in the Military Domain (REAIM) Regional Consultations for Asia on 26-27 February 2024.[66]

By contrast to their official views on AI in defence, concrete defence acquisitions of AI-enabled systems can be more broadly seen as a reflection of specific states' threat perceptions and interests. As summed up in Figure 1 above, several ASEAN states seek the technology to enhance their capabilities for enhanced mission command, intelligence and surveillance, protection and fires capabilities. These seem to be geared toward meeting territorial defence challenges: the Philippines is embroiled in intense maritime and territorial disputes with China over features and maritime jurisdictions in the South China Sea (SCS) and is sensitive to both overt military and non-military coercive measures, including attempts to steal its data. Malaysia also has maritime and territorial claims in the SCS. While not directly affected by the disputes, Indonesia and Singapore are affected by tensions in the SCS, harbour growing concerns on the regional balance, and have critical and/or expansive waterways and airspace to monitor and protect. Thailand's drive to defence capability development may be influenced by a general need to maintain defence capacity, which was subsequently proven given the resumption of a long-running border conflict with Cambodia throughout 2025.

However, the piecemeal implementation of AI among Southeast Asian militaries, and in mostly niche capabilities compared to the expansive programs of the United States and China, demonstrates differences in doctrinal focus and approach, in addition to more familiar issues of funding, and the abovementioned perceived need for caution and further study expressed in Track II fora and the ADMM Joint Statement on AI in the Defence Sector.

---

[63] List of Participants, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 10 September 2025, available 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-INF.1.pdf.

[64] Draft Protocol VI Submitted by Argentina, Ecuador, Costa Rica, El Salvador, Guatemala, Kazakhstan, Nigeria, Panama, the Philippines, Sierra Leone and Uruguay, 9 August 2022, available at https://docs-library.unoda.org/Convention_on_Certain_Conventional_ Weapons_-_Group_of_Governmental_Experts_(2022)/CCW-GGE.1-2022-WP.8.pdf.

[65] Inputs from the Philippines, United Nations Secretary General Report on Resolution 78/241 (Lethal autonomous weapon systems), 17 May 2024, p. 1, available at: https://docs-library.unoda.org/General_Assembly_First_Committee_-Seventy-Ninth_session_(2024)/ 78-241-Philippines-EN.pdf.

[66] Singapore's National Submission on the Topic of Lethal Autonomous Weapon Systems, 9 May 2024, p. 3, available at: https://docs-library.unoda.org/General_Assembly_First _Committee_-Seventy-Ninth_session_(2024)/78-241-Singapore-EN.pdf.

Challenges for ASEAN states' adoption of AI

A careful reading of the ADMM Joint Statement may indicate that ADMM's understanding of AI remains largely cybersecurity focused, evident by their intention to develop knowledge and capacities for cooperation on AI under the ambit of ASEAN's cybersecurity platforms.[67] This, as well as the abovementioned interest in non-warfighting applications of AI in defence, may be explained by ADMM's general reticence to discuss military hardware and platforms (which would include autonomous systems such as LAWS). This is perhaps in keeping with ASEAN norms on non-interference in members' affairs[68], considering that the subject of procurements could quickly become sensitive issues due to the still extant territorial disputes between members.

It should also be noted that while ADMM has a growing portfolio of defence cooperation initiatives; ASEAN itself is not a security alliance in the same way as the North Atlantic Treaty Organization (NATO). Its members are unlikely to form such a similar defence pact due to divergent defence and security interests, as acknowledged publicly by at least one ASEAN member state's defence secretary.[69] This is before taking into account actual conflict between member states, as exemplified most recently by the Cambodia-Thailand border clashes. By comparison, ADMM have had extensive discussions on cybersecurity and cyber defence, established an Expert Working Group (EWG) on Cyber Security in 2017, the ADMM Cybersecurity and Information Centre of Excellence (ACICE) and the ASEAN Cyber Defence Network in 2021.[70]

Using cybersecurity as a base for AI cooperation may be valuable in protecting generative AI and LLMs, especially if the AI and their databases are locally hosted. Local hosting and isolating the training databases would help reduce cyber risks,[71] compared to just using models and databases with regular access to the open internet, which may open them to the risk of incorporating disinformation that can negatively impact the assessments AIs provide.[72]

---

[67] ADMM, above note 60, p. 3.
[68] Treaty of Amity and Cooperation in Southeast Asia, 24 February 1976, Art. II.
[69] Martin Sadongdong, "Is ASEAN ready to have a NATO-like counterpart? 'Not possible at this time', says Teodoro", *Manila Bulletin*, 5 November 2024, available at: https://mb. com.ph/2024/11/5/asean-nato-counterpart-teodoro.
[70] ASEAN Cybersecurity and Information Centre of Excellence Update on The Cyber Domain, March 2025, p. 5 available at: https://www.acice-asean.org/files/cybersecurity %20centre%20reports/mar25_cyber.pdf.
[71] Daniel Anderson, "How to Run AI Locally: Benefits, Challenges, and Solutions", *Autonomous*, 29 July 2025, available at: https://www.autonomous.ai/ourblog/run-ai-locally.
[72] Mackenzie Sadeghi, "Top 10 Generative AI Models Mimic Russian Disinformation Claims A Third of the Time, Citing Moscow-Created Fake Local News Sites as Authoritative Sources", *NewsGuard*, 18 June 2024, available at: https://www.newsguardtech.com/special -reports/generative-ai-models-mimic-russian-disinformation-cite-fake-news/.

Protecting networks is essential to effective use of AI LLMs especially if these systems will be used to inform operational planning and ultimately targeting decisions, akin to Israel's *Habsora* and the United States' Maven. The reliability of current AI applications for operational planning is debatable even without the risk of data tampering or flooding of malicious data, due to the likelihood that an AI's training database may have plenty of information on what constitutes targets, but not enough information on what does not constitute targets, as was reportedly found out in initial Israeli trials of what would become *Habsora* in 2021.[73]

Cybersecurity and data protection may be even more paramount if off-the-shelf commercial AI models are used for such purposes. Costs could be reduced by the adoption of these commercial models for intelligence consolidation, analysis and assistance to mission command, in comparison to having to devise bespoke models suitable for Southeast Asian states' needs, and this is a tempting prospect due to the general need for Southeast Asian states to balance their defence spending with other development priorities. However, relying on off-the-shelf models come with their own risks. In general, commercial AI LLM models pose an inherent level of data security risk, due to their dependence on databases maintained on the open web, and usually in data centres that may not be within the user states' borders.

There is also the basic issue of digital sovereignty; the leading LLM models are either developed by firms in the United States or China,[74] which opens up concerns regarding data security and control. Some companies, such as Anthropic, have expressed concerns, if not total opposition, to certain military uses of their Claude AI, which may have been used in assisting the planning of Operation Absolute Resolve that led to the United States' capture of Venezuelan president Nicolas Maduro in January 2026.[75] Regardless of where one stands on Anthropic's position, a private firm deciding on what uses are permissible for its AI will likely influence some states' decisions to use their services. That same AI was also recently the subject of a major cyber-attack, where hackers of alleged Chinese origins were able to subvert Claude to conduct cyber-attacks on a number of corporations and government bodies,[76] reinforcing the security risks of using commercial AI.

Finally, there is the need to minimize the data potentially accessible to outsider and adversary AI; the Philippine 2023 internal memo banning use of certain

---

[73] Jewish Institute for the National Security of America, *Gaza Conflict 2021 Assessment: Observations and Lessons*, Washington D.C., October 2021, p. 31, available at: https://jinsa.org/wp-content/uploads/2021/10/Gaza-Assessment.v8-1.pdf.

[74] "AI Leaderboards", *LLMStats*, 11 March 2026, available at: https://llm-stats.com.

[75] Paulo Carvão, "Should AI Go To War? Anthropic And The Pentagon Fight It Out", *Forbes*, 20 February 2026, available at: https://www.forbes.com/sites/paulocarvao/2026/02/20/ should-ai-go-to-war-anthropic-and-the-pentagon-fight-it-out/.

[76] David Klepper and Matt O'Brien, "Anthropic warns of AI-driven hacking campaign linked to China", *Associated Press*, 15 November 2025, available at: https://apnews. com/article/ai-cyber-china-hacking-artificial-intelligence-anthropic-4e7e5b1a7df946169c 72c1df58f90295.

AI apps is one example of a countermeasure intended to deny or at least hobble potential sources of sensitive personal information for adversaries.

Limitations of AI: not a panacea for lack in military capability

To what extent AI can grant advantages to smaller states which are not necessarily leaders in the technology, or lack the necessary infrastructure to fully realize aspirations of a self-reliant or independent system? While AI is a disruptive technology that could grant asymmetric advantages for states that invest in them, there is the distinct possibility that AI by itself will not be sufficient to gain reverse asymmetry[77] for small states, wherein the smaller state aims to leverage military innovation to overcome larger opponents. This is due to the reality that great powers not only have more resources than smaller states, but are also often technology leaders with first-mover advantages and at least a temporary monopoly over new military technologies. [78]

For AI to be able to help a state win, that state will still need sufficient quantities of everything from munitions to sensors, (region-specific) training data, secure supply chains and motivated and trained soldiers. This can be seen in the ongoing Russian invasion of Ukraine; while the adoption of unmanned systems is moving rapidly in Ukraine's armed forces, including successful field tests of a variety of unmanned systems[79], the outcome of the war remains uncertain, in part due to the Russians' own adaptability and still-significant conventional and numerical preponderance. Even as emerging technologies are implemented, war is ultimately still a contest of wills[80] and attrition in materiel and lives. This is made worse by the fact that, unlike most other technologies identified as providing asymmetric advantage such as drones, AI is not necessarily cheap to invest in[81], requiring substantial investments in data centre infrastructure. Quite apart from the up-front capital costs, AI data centre infrastructure will likely pose additional strains on scarce

[77] Michael Raska, "Chapter 3 - Creating a reverse asymmetry: military innovation concepts, issues and debates in the IDF", in *Military Innovation in Small States: Creating a reverse asymmetry*, Routledge, London & New York, 2016, p. 91.
[78] Francis Domingo, "Distribution of Power and Cyber Capability Development", in *Making Sense of Cyber Capabilities for Small States: Case Studies from the Asia-Pacific*, Routledge, London, 2022, p. 96.
[79] Andrew White, "Why Ukraine's all-drone, multi-domain attack could be a 'seminal' moment in warfare", *Breaking Defense*, 24 January 2025, available at: https:// breakingdefense.com/2025/01/why-ukraines-all-drone-multi-domain-attack-could-be-a-seminal-moment-in-warfare/.
[80] Michael Raska and Richard A. Bitzinger, "Introduction: The AI Wave in Defence Innovation", in *The AI Wave in Defence Innovation - Assessing Military Artificial Intelligence Strategies, Capabilities and Trajectories*, Routledge, London & New York, 2023, p. 9.
[81] National Defense College of the Philippines, "Defense industrial ecosystem dev't highlighted in first Security Lounge for 2026", 27 February 2026, available at: https:// ndcp.edu.ph/defense-industrial-ecosystem-devt-highlighted-in-first-security-lounge-for-2026/.

water and energy resources,[82] which are already major concerns for several Southeast Asian states.

Expectations that AI can give asymmetric advantage must be weighed against the physical and material readiness of Southeast Asian military forces. It should be noted that even as they are modernizing their forces, they still do not yet have the preponderance of force to achieve their desired force objectives, such as establishing an anti-access/area-denial (A2/AD) network.[83]

Ways Forward

AI in defence is still very much an evolving and growing area of development and study. Even as militaries seek the speed and efficiency that AI promises, there is reason for a more cautious approach. The alleged impact of decision-support AI in Gaza points to the potential collateral damage and casualties that could happen if speed of generating targets is prioritized above all other considerations.[84]

As the decision to attack still ultimately lies with the analysts and commanders who recommend and make the decisions to attack, this implies that future use of such decision-support AIs must be coupled with increased training requirements for future operators,[85] aiming to increase awareness of the essential principles on the law of armed conflict (LOAC) and international humanitarian law (IHL). While unlikely to be foolproof, it must be ensured that the officers that sign off on AI-recommended attack orders know clearly the risks of blindly trusting AI systems and their recommendations.

Military forces may opt to design specific military directives most suitable for providing specific guidance and delimitation for AI systems; be it rules of engagement (ROE), fragmentary orders (FRAGO), operational orders (OPORD) or other appropriate instructions. The skill and competence to design such instructions may need to be part of future training requirements, wherein the desire for

---

[82] Paul Barry, *2025-2026 Data Centre Construction Cost Index*, Turner & Townsend, 5 November 2025, available at: https://reports.turnerandtownsend.com/data-centre-construction-cost-index-2025/.

[83] Evan Laksamana, "6. Implications and Considerations", *Contingent Capabilities: Southeast Asia's Emerging Anti-Access Environment*, International Institute of Strategic Studies, London, February 2026, p. 33, available at: https://www.iiss.org/globalassets/ media-library---content--migration/files/research-papers/2026/contingent-capabilities-southeast-asias-emerging-anti-access-environment/contingent-capabilities-southeast_ asias-emerging-anti-access-environment.pdf.

[84] N. Sylvia, above note 24.

[85] Erick Nielson C. Javier, *Reality from Science Fiction: Strategic Implications of Autonomous Weapon Systems for Philippine Defense and Deterrence*, NDCP Executive Policy Brief No. 2024-01, National Defense College of the Philippines, Quezon City, 27 March 2024, p. 6, available at: https://ndcp.edu.ph/wp-content/uploads/2024/04/EPB-2024-01.pdf.

precautions on use of AI must be balanced with operational suitability and technical competence.[86]

At the ASEAN level, proposals to improve ADMM's cooperation on military AI currently revolve around areas of practical cooperation, particularly non-combat operations such as maritime security exercises, as well as exercises for responding to cyber threats to AI systems.[87] These are in line with ASEAN's preferred approach for cooperation, meant to build confidence among members. Given that the largest benefits and highest risks of military AI tend to involve high decision-making speeds, it is likely that stress-testing such AI, which is essential to really establish actual baselines on AI models, will need different operational scenarios, specific to the ASEAN region, to better visualize the challenges and thresholds that future commanders must keep in mind when using AI for decision support. While exercising use of military AI in high-intensity scenarios at the ASEAN level may be politically unlikely at present, moving towards such exercises has to be considered sooner or later; otherwise, individual ASEAN member states may opt to investigate building such capabilities in partnership with other willing members, or with third party states. While building with third-party state providers is expedient and provides access to capabilities probably not available yet within ASEAN, it creates potential strategic vulnerabilities that ASEAN states would rather avoid.

In terms of AI-assisted weapons and materiel, there seem to be no clear consensus yet amongst ASEAN member states on standards and practices. ASEAN and ADMM could build on the Joint Statement of the ASEAN Defence Ministers on Cooperation in the Field of Artificial Intelligence in the Defence Sector and explore discussing guidelines, standards and practices for material applications of AI; perhaps as a means of strengthening ASEAN Defence Industry Collaboration (ADIC)[88], another ongoing ADMM initiative for concrete defence cooperation.

Ultimately, while AI is likely to derive some marginal efficiencies in force command and control as seen in initial experiments, or provide some niche operational capability, much will need to be done before AI can result in enhanced effectiveness for Southeast Asian militaries, given those states' other capability gaps. Southeast Asian states would face other factors that may mitigate whatever advantage AI could bring, such as rules of engagement considerations or more

---

[86] Jonathan Kwik *et al.*, *Controlling Military Artificial Intelligence: Harnessing Rules of Engagement and Military Directives*, Asser Policy Brief 2025-01, TM Asser Institute, The Hague, February 2025, pp. 21-22, available at: https://papers.ssrn.com/sol3/papers.cfm ?abstract_id=5132731.

[87] Mohammed Faizal Bin Abdul Rahman, *AI in Defence: The Way Forward for ASEAN's ADMM Cooperation*, IDSS Paper No. 084, S. Rajaratnam School of International Studies, Singapore, pp. 3-4, available at: https://www.rsis.edu.sg/wp-content/uploads/2025/08/ IP25084.pdf.

[88] Concept Paper on Enhancing ASEAN Defence Industry Collaboration adopted by the 19th ASEAN Defence Ministers Meeting, 31 October 2025, available at: https://admm.asean.org/dmdocuments/2025_19th%20ADMM_KL,%20Malaysia,%2031%20October%202025_Concept%20Paper%20on%20Enhancing%20ADIC.pdf.

fundamental disparities in arms and capabilities. Like other countries, the key for Southeast Asian states to enhance their security will be to combine the adoption of AI with the requisite material capabilities, and strategic, organizational and operational adaptability, that mark effective military forces.[89]

The use of AI in the military will likely only increase, especially as the technology matures and is successfully demonstrated in other regions of the world. As the militaries of Southeast Asia prepare to meet a future of warfare where AI plays major roles, they must be ready to balance the imperative to experiment with prudence and caution.

[89] M. Raska and R. Bitzinger, above note 81, p. 9.

Quo Vadis International Humanitarian Law in AI-Enabled Conflicts: A Southeast Asia Regional Perspective

Bagus Jatmiko[*]
Indonesian Naval Command and Staff College

ABSTRACT

AI-enabled autonomous weapons promise operational precision but introduce systemic risks — algorithmic bias, black-box opacity, reduced human control, and value misalignment (including sycophantic tendencies) — that undermine jus in bello principles of distinction, proportionality, and precaution. In Southeast Asia, these flaws risk escalating conflict and causing disproportionate harm across its diverse, densely populated maritime domain, while Great-Power AI dominance constrains Middle-Power agency. Using a qualitative, legal-normative approach, this paper argues that without binding international norms, AI's inherent flaws will erode IHL compliance. Urgent multilateral action through CCW protocols and the REAIM framework, supported by ASEAN-led regional governance, is essential.

Keywords: Autonomous Systems, Biased AI, IHL Compliance, Jus in Bello, CCW Protocols, REAIM Framework, ASEAN Perspective

1. Introduction

The widespread use of artificial intelligence (AI) in the military domain has significantly transformed how the military conducts its operations. Countries, particularly the Major Military Powers, are steadily integrating AI-enabled technologies into their military capabilities, including, but not limited to, intelligence gathering and analysis processes, AI-decision support systems (AI-DSS), logistics procedures, cyber operations, and, most controversially, the deployment of autonomous, semi-autonomous, and lethal autonomous weapons systems (LAWS), to conduct hostilities during an open conflict or when it escalates to war.[1]

---

[*] Bagus Jatmiko holds a Doctoral Degree in Information Sciences from the Naval Postgraduate School in Monterey, California. His professional background is as a seasoned officer in the Indonesian Navy, where he holds the rank of captain and currently serves as the head of the Wargaming Centre at the Indonesian Naval and Command Staff College.
In the field, he served as a field operative and staff officer at various levels of the military, particularly the Navy. His strong academic and professional experience provides a seamless blend of perspectives on AI implementation in the military domain, and he is determined to ardently voice the Global South viewpoint.
E-mail preference: Gusmiko4701@gmail.com

[1] Euysun Hwang, "Lethal Autonomous Weapons: The next Frontier in International Security and Arms Control", *Stanford International Policy Review*, 30 January 2025, available at:

On the one hand, proponents of AI argue that the further implementation of AI systems into military technologies may, in fact, improve compliance with International Humanitarian Law (IHL), which embodies jus in bello principles. Compliance is achieved through increased operational efficiency, reduced human-made mistakes, faster cycles, more accurate decision-making, refined precision targeting that enables accurate proportionality assessments, improved target discrimination, and reduced risks to military personnel.[2] In the view of some, AI-enabled weapons could be the harbinger of the next military revolution.[3] On the other hand, such claims remain contested in global discourse, particularly when an AI system relies on an unreliable, opaque deep-learning architecture whose reasoning and processing are not fully explicable or auditable, forming an incomprehensible "black box."[4]

During the informal consultation on LAWS held by the United Nations (UN) in New York on 12 May 2025, Secretary-General António Guterres described autonomous weapon systems as "politically unacceptable and morally repugnant," particularly those that completely remove human intervention from decision-making processes, or the human-in-the-loop concept. His strong statement reflects the growing global concern over the uncontrolled proliferation and deployment of autonomous weapon systems and their implications for how conflicts are fought.[5] Furthermore, the global debates revolve around the legality and regulation of LAWS within the framework of the UN Convention on Certain Conventional Weapons

---

https://fsi.stanford.edu/sipr/content/lethal-autonomous-weapons-next-frontier-international-security-and-arms-control (all internet references were accessed on 3 July 2026); Anthony King, "Digital Targeting: Artificial Intelligence, Data, and Military Intelligence", *Journal of Global Security Studies*, Vol. 9, No. 2, 2024; Priyesh Mishra et al., "Code, Command, and Conflict: Charting the Future of Military AI", *Belfer Center for Science and International Affairs*, 2025, available at: https://www.belfercenter.org/sites/default/files/2025-12/Belfer_Code%20Command%20and%20Conflict_1.1.pdf; Kurtis H. Simpson et al., "Militarizing AI: How to Catch the Digital Dragon?", *Centre for International Governance Innovation*, 26 February 2025, available at: https://www. cigionline.org/articles/militarizing-ai-how-to-catch-the-digital-dragon/.

[2] Greg Allen and Taniel Chan, "Artificial Intelligence and National Security", *Belfer Center Study*, 2017; Ronald Arkin, *Governing Lethal Behavior in Autonomous Robots*, CRC Press, Boca Raton, 2009; Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, W.W. Norton, New York, 2018.

[3] Denise Garcia, *The AI Military Race: Common Good Governance in the Age of Artificial Intelligence*, Oxford University Press, Oxford, 2024.

[4] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms", *Big Data & Society*, Vol.3, No.1, June 2016; Roy Lindelauf and Herwin Meerveld, "Building Trust in Military AI Starts with Opening the Black Box", *War on the Rocks*, 12 August 2025, available at: https://warontherocks.com/2025/08/building-trust-in-military-ai-starts-with-opening-the-black-box/.

[5] UN Press Release, *Lethal Autonomous Weapon System 'Politically Unacceptable, Morally Repugnant and Should Be Banned', Secretary-General Says during Informal Consultations on Issue*, UN Meetings Coverage and Press Releases, New York, 12 May 2025, available at: https://press.un.org/en/2025/sgsm22643.doc.htm.

(CCW), while acknowledging their potential to provide significant advantages over military operations and, at the same time, pose profound humanitarian risks.[6]

From a technically agnostic perspective, AI-enabled systems in warfare offer capabilities to analyse anomalies and assess vast amounts of data and information, significantly faster than human operators. Hence, they enable militaries to respond rapidly to imminent threats in a contested environment. Nevertheless, these technologies are not without flaws. They face constraints that limit reliability and adaptability, particularly in legal and regulatory matters.[7] These gaps are sources of debate on AI implementation, particularly in the military domain. The technologies discussed encompass machine-learning-based intelligence, surveillance, and reconnaissance (ISR) capabilities; predictive and computational analytics tools for planning and operational purposes; and platforms that can be deployed on the battlefield to identify and engage targets using input data and programmed algorithms.

In general, despite their operational appeal, the implementation of AI-enabled military technologies during war is characterized by several structural vulnerabilities that challenge the central role of humans in *jus in bello*, thereby potentially undermining proper IHL implementation and distinct accountability. This article identifies four inherent flaws of AI. The first is algorithmic bias that could lead to systematic misinterpretations in the identification of legitimate targets. The second is the dilemma of black boxes, defined as the limitations of transparency and explainability in the algorithm's decision-making process. Third is the significant or complete elimination of human control in autonomous systems, which jeopardizes accountability. The final one is the emergence of misalignment behavior, which could result from the black-box dilemma, leading systems to provide inexplicable suggestions that may disregard existing norms and legal principles for operational gains.

This paper pursues three interrelated objectives. It first examines how the normative framework of *jus in bello* may be disrupted by the characteristics of AI systems, in particular bias, opacity, autonomy, and value misalignment. Second, it examines how these disruptions may heighten vulnerability for populations in regions such as Southeast Asia, given their archipelagic characteristics, densely populated coastal zones, and propensity for maritime conflicts. Third, it assesses the feasibility and capacity of multilateral and regional governance initiatives, including discussions under the CCW, the Responsible AI in the Military Domain (REAIM), and the Association of Southeast Asian Nations (ASEAN) cooperation frameworks and approaches, to align the proliferation of AI systems with IHL objectives and principles.

---

[6] CCW Group of Governmental Experts, *Report of the 2023 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, CCW/GGE.1/2023/CRP.2, United Nations, Geneva, 2023.
[7] E. Hwang, above note 1.

The implementation of LAWS in strategic and humanitarian contexts of Southeast Asia has received relatively little attention, despite previous research that has thoroughly examined LAWS from global or Western-centric perspectives.[8] The region has distinct characteristics that pose challenges for the deployment and assessment of AI-based weapon systems, including contested maritime domains, intricate sovereignty disputes, and dense civilian populations living near military installations and shipping lines.[9] AI-enabled errors or biased targeting might quickly escalate conflicts and cause disproportionate humanitarian harm in such asymmetrical and archipelagic situations.

Regarding geopolitical positioning, Southeast Asian countries balance their reliance on technology, strategic alignment, and normative commitments to ASEAN centrality, often occupying intermediary positions between the Major Powers. Because of the region's susceptibility to both technology diffusion and geopolitical conflict, context-specific governance frameworks that supplement global IHL standards are essential. This paper contributes to ongoing discussions on the future direction—*Quo Vadis?*—of IHL in the era of military artificial intelligence, focusing on the time of conflict, by tying international legal discussions to a Southeast Asian regional perspective.

The paper proceeds by elaborating the research methodology, followed by a comprehensive review of the literature on military AI applications, associated ethical concerns, and evolving interpretations of international humanitarian law (IHL). It then presents a theoretical framework integrating legal, normative, and regional security perspectives, examines the impact of AI's inherent deficiencies on *jus in bello* principles, evaluates key global governance initiatives including the CCW negotiations and REAIM process, analyses the Southeast Asian context with particular attention to ASEAN-led frameworks and norm-building efforts, synthesizes the findings while addressing counterarguments, offers policy recommendations, and concludes with reflections on the future of human-centric IHL amid the accelerating militarization of artificial intelligence.

[8] Magdalena Pacholska, "Military Artificial Intelligence and the Principle of Distinction: A State Responsibility Perspective", *Israel Law Review*, Vol. 56, No. 1, 2023.
[9] Srishti Chhaya, "The $5.3 Trillion Question: How South China Sea Tensions Are Rewriting Global Trade Rules", *Atlas Institute for International Affairs*, 4 July 2025, available at: https://atlasinstitute.org/the-5-3-trillion-question-how-south-china-sea-tensions-are-rewriting-global-trade-rules/; Abdussalam Giuma A. Triki, "Understanding the South China Sea Crisis: State Claims, International Interventions, and Implications", *Asia & the Pacific Policy Studies*, Vol. 13, No. 1, 2026.

2. Research Methodology

For this study, the methodology is a qualitative, conceptual-analytical paper that employs legal-normative analysis and a secondary-source review. This approach is appropriate for explaining the nature of the research problem: the erosion of *jus in bello* principles by AI's inherent flaws, which may be embedded in existing and future autonomous weapon systems. This Paper constructs and defends a scholarly argument by systematically analyzing the relevant legal framework, the technical characteristics of AI systems, and the structural features of Southeast Asia's regional security environment, drawing on the best available secondary sources.

The methodology chosen for this paper rests on three interlocking justifications — epistemological, empirical, and practical — each of which responds directly to the nature of the research problem. Epistemologically, this study is firmly grounded in an interpretive tradition to accommodate the research question of whether AI's vulnerabilities undermine the principles of IHL and what governance frameworks can address this issue, which are inherently normative and interpretive. International humanitarian law, the central institution under examination, is not a fixed object with stable, observer-independent properties; it is a social and legal construction whose meaning, applicability, and enforceability are continuously negotiated among States, international organizations, non-governmental actors, and legal scholars. This interpretivist orientation is consistent with—and, in fact, necessitates —the adoption of constructivism as the paper's primary theoretical framework, as detailed in the conceptual framework section below.

Moreover, the empirical landscape of AI-enabled autonomous weapons in armed conflict is currently characterized by significant data inaccessibility. States that deploy AI-assisted targeting systems do not publish operational performance data in forms accessible to independent researchers. Incidents involving the broader use of automated decision-support in contemporary conflicts are documented primarily through investigative journalism and NGO reporting[10] rather than through systematic empirical datasets amenable to quantitative analysis. The absence of a reliable empirical dataset is not a methodological shortcoming unique to this paper; it is a structural feature of the research environment that any scholarly approach must acknowledge and deal with. Within these constraints, the conceptual-analytical method extracts maximum analytical value from available secondary evidence —

[10] Human Rights Watch, "Questions and Answers: Israeli Military's Use of Digital Tools in Gaza", *Human Rights Watch*, 10 September 2024, available at: https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-digital-tools-gaza., *Human Rights Watch*, 2024.

including technical assessments of AI bias,[11] legal doctrinal analyses[12], governance assessments[13], and regional security analyses[14] — by subjecting these sources to structured theoretical interpretation rather than treating them as isolated data points.

From a practical perspective, the paper is situated at the intersection of international law, security studies, and regional political analysis—a multidisciplinary space in which conceptual-analytical methodologies have a well-established and respected tradition. Hence, the contribution lies precisely at this disciplinary intersection—and the conceptual-analytical method is the approach best suited to bridge the legal, strategic, and regional dimensions of the AI-IHL problem. Nevertheless, this methodological choice carries acknowledged limitations. The paper cannot claim to demonstrate empirically that AI systems have in fact violated IHL in any specific instance, nor can it measure the precise magnitude of AI-attributable risk. It advances interpretive and normative claims that are subject to scholarly contestation and revision as the empirical and normative landscape evolves. These limitations are inherent to the research domain and are shared by all serious scholarship in this area. They underscore, rather than undermine, the need for continued interdisciplinary engagement of exactly the kind this paper undertakes.

Henceforth, the conceptual-analytical method is suitable for examining situations in which the phenomenon of interest (AI-enabled autonomous weapons) is considered too novel, rapidly evolving, and empirically inaccessible to support systematic quantitative research, yet too consequential to wait for long-term data accumulation before scholarly and policy communities engage. Comparative and analytical methods in this mode are recognized as rigorous contributions to knowledge when their evidential base, analytical logic, and interpretive claims are made explicit and subject to critical scrutiny — which this methodology section aims to ensure.

---

[11] Laura Bruun and Marta Bo, *Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law*, Stockholm International Peace Research Institute, Stockholm, 2025; Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 2018.

[12] Jonathan Kwik and Tom Van Engers, "Algorithmic Fog of War: When Lack of Transparency Violates the Law of Armed Conflict", *Journal of Future Robot Life*, Vol. 2, No. 1-2, Amsterdam, June 2021; Elliot Winter, "The Compatibility of Autonomous Weapons with the Principles of International Humanitarian Law", *Journal of Conflict and Security Law*, Vol. 27, No. 1, 2022.

[13] Amna Batool, Didar Zowghi and Muneera Bano, "AI Governance: A Systematic Literature Review", *AI and Ethics*, Vol. 5, No. 3, 2025; Herwin Meerveld et al., "Operationalising Responsible AI in the Military Domain: A Context-Specific Assessment", *Ethics and Information Technology*, Vol. 27, No. 4, 2025.

[14] Muhammad Jordan Baresi, Anak Agung Banyu Perwita and Yermia Hendarwoto, "U.S.-China Rivalry Controls AI-Based Defense Machine in the Natuna Sea and Malacca Strait: Indonesia's AI Governance Diplomacy in Navigating International Regulations", *Asian Journal of Science, Technology, Engineering, and Art*, Vol. 3, No. 5, 2025; Bama Andika Putra, "Governing AI in Southeast Asia: ASEAN's Way Forward", *Frontiers in Artificial Intelligence*, Vol. 7, 2024.

3. Literature Review

Numerous important studies on the subject are summarized in this paper to support its overall arguments. Bias in training data can result in discriminatory targeting, particularly against diverse groups. Further, it exacerbates problems like algorithmic bias, opacity, and accountability gaps. However, the effort to achieve explainability and oversight, to lessen the perceived problem, is seriously hindered by the presence of black-box characteristics in the systems. [15] Moreover, research on anthropomorphism and automation bias reveals the dangers of excessive reliance on AI, which could compromise human moral agency in critical, decisive judgment that often occurs during war. [16] Legal critiques highlight problems with attributing violations and advocate transparency and moral design to reduce impunity.[17] During times of conflict, these shortcomings raise questions about proportionality and prudence, as AI may not be able to interpret human values. Furthermore, non-Western viewpoints are not fully integrated into the existing literature, which primarily draws on Western ethical frameworks and contexts.[18]

Current research is particularly concerned about challenges posed by AI to the fundamental IHL principles outlined in the Geneva Conventions and Additional Protocols — such as distinction, proportionality, and precaution. Nevertheless, AI's opacity and unreliability could compromise these principles, as systems struggle to make the contextual judgments required for legitimate attacks.[19] While recognizing shortcomings in accountability for unintentional harm, UN documents and working papers emphasize the need for human-centric measures to ensure compliance with IHL and regulations.[20] This compliance becomes particularly relevant, as studies focusing on *jus in bello* highlight that AI has the potential to undermine human agency, increasing the risks of initiating conflict or even war by lowering the threshold for conflict.[21] Additionally, to address technological disruptions, critics contend that current frameworks are inadequate for LAWS and call for new norms and standards.[22]

---

[15] L. Bruun and M. Bo, above note 11.
[16] Adriana Placani, "Anthropomorphism in AI: Hype and Fallacy", *AI and Ethics*, Vol. 4, No. 3, 2024.
[17] Erika Feller, *Impunity, Accountability and Respect for Human Rights and the Rule of Law*, Issues Brief, The Initiative for Peacebuilding, The University of Melbourne, 2024; ICRC, *Investigating and Prosecuting Serious Violations: An Important Tool against Impunity*, International Committee of the Red Cross, 14 October 2024.
[18] M. Pacholska, above note 8.
[19] ICRC, *Investigating and Prosecuting Serious Violations,* above note 17; M. Pacholska, above note 8.
[20] 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*, Vol. 27, No. 4, 2025.
[21] Francis Grimal and Michael J. Pollard, "A Blurring of Lines: The *Jus Ad Bellum* Past, Present, and Future", *Journal on the Use of Force and International Law*, Vol. 11, No. 1-2, 2024.
[22] Jeroen K.G. Hopster and Matthijs M. Maas, "The Technology Triad: Disruptive AI, Regulatory Gaps and Value Change", *AI and Ethics*, Vol. 4, No. 4, 2024.

On a broader level, the current global AI standards are shaped by investments and strategic conflicts between Major Powers, especially the US and China, which control the development of AI technologies at scale. The literature outlines Middle Powers' methods, including bandwagoning or independent paths, by juxtaposing US innovation with China's state-driven adoption.[23] Amid US-China competition, there are calls for specific frameworks to address this geostrategic development, emphasizing ASEAN's role in maritime security and in ethical AI governance in the Indo-Pacific.[24] Moreover, research found Southeast Asia to be particularly susceptible to conflict due to the absence of a regional AI governance framework with binding power to address relevant challenges, particularly in the security and defence (military) domain.[25]

Furthermore, significant gaps remain despite the expanding body of work, especially in Southeast Asian-specific assessments. The majority of research overlooks the specificity of archipelagic and predominant maritime domain vulnerabilities, combined with diverse and dense populations and asymmetric conflicts in the Indo-Pacific, in favor of Western or global contexts. Non-Western values are frequently disregarded in the literature on ethical AI, and there is little examination of cultural diversity in governance frameworks. Both the integration of AI ethics with maritime security and the regional disparities issues within ASEAN are understudied. Additionally, there are a few empirical studies on AI's humanitarian effects in simulations from Southeast Asia, but a dearth of studies on Middle- and Small-Power AI strategies in the face of Great-Power supremacy. These gaps highlight the need for contextually grounded research to address the specific challenges posed by AI-enabled conflict in Southeast Asia.

4. Conceptual Framework

The primary conceptual framework adopted in this paper is constructivism, which holds that international norms and rules are socially constructed through interactions among States, institutions, and non-state actors, and that they develop through discourse and common understandings. [26] Constructivism sheds light on how shortcomings such as bias and opacity undermine and transform the human-centric principles of IHL in the context of AI in combat, thereby driving international norm-building in forums such as the CCW, the REAIM initiative, and the regional ASEAN AI framework. This theory supports the paper's argument that precautionary duties grounded in ethical design, transparency, and human control—developed through international discussions—are required in light of AI's disruptions to *jus in bello*, such as the erosion of distinctions and proportionality. Instead of

[23] Francisco Javier Varela Sandoval et al., "How Middle Powers Can Weather US and Chinese AI Dominance", *Chatham House*, London, February 2026.
[24] M.J. Baresi *et al.*, above note 14.
[25] B.A. Putra, above note 14.
[26] Michael L. Barnett, "Constructivism", in Alexandra Gheciu amd William C. Wohlforth (eds), *The Oxford Handbook of International Security*, Oxford University Press, 2018.

treating laws as static, constructivism emphasizes the interpretive adaptation of current IHL to new technologies. It supports arguments for human-centered strategies that preserve moral agency in AI-supported decisions. Moreover, this theory supports the call for legally enforced protocols to mitigate AI biases and preserve alignment with humanitarian values in asymmetric conflicts by presenting norms as socially constructed among the involved actors.

To complement constructivism, realism is a supporting theory that emphasizes the influence of power interests, national security, and rivalry on state conduct in international affairs, especially in technology sectors such as artificial intelligence.[27] Realism highlights how Major Powers like the US and China control the development of military AI, prioritizing competitive advantage over moral considerations, let alone compliance with international ethics and norms. This approach influences middle- and small-state decisions about whether to bandwagon or pursue their own path. Realist interests may jeopardize multilateral endeavours like the CCW, but they also offer weaker States opportunities to shape outcomes through alliances.[28] This theory aligns with constructivism by emphasizing power imbalances in the development of norms. Thus, realism puts the perils of AI bias, which preserves impunity, into perspective. These precedents are highly likely because Major Powers are expected to oppose legally mandated standards that curtail their technical advantage, thereby increasing humanitarian vulnerabilities in regions with predominantly Middle- and Small-Power States, such as Southeast Asia.[29]

By examining ASEAN as a norm-entrepreneur in regional security and promoting consensus-based governance amid Great-Power rivalry, the framework applies these theories to Southeast Asia and highlights ASEAN's significance. Archipelagic state theory informs this regional perspective, which highlights vulnerabilities in maritime disputes (e.g., the South China Sea conflict),[30] where misaligned AI could intensify asymmetric conflicts and harm diverse civilian populations. While realism emphasizes the necessity of well-rounded tactics to offset the effect of Dominating Powers, constructivism promotes ASEAN's role in creating

---

[27] Lavinia-Maria Savu, "Realism, Liberalism and Constructivism in the Pursuit of Security", *International Scientific Conference 'Strategies XXI'*, Bucharest, Vol. 17, 2021.

[28] Ingvild Bode, "Norm-making and the Global South: Attempts to Regulate Lethal Autonomous Weapons Systems", *Global Policy*, Vol. 10, No. 3, September 2019; Anna Nadibaidze, "Great Power Identity in Russia's Position on Autonomous Weapons Systems", *Contemporary Security Policy*, Vol. 43, No. 3, 2022; Goddy Uwa Osimen *et al.*, "Artificial Intelligence and Arms Control in Modern Warfare", *Cogent Social Sciences*, Vol. 10, No. 1, 2024.

[29] Jie Guo, "The Ethical Legitimacy of Autonomous Weapons Systems: Reconfiguring War Accountability in the Age of Artificial Intelligence", *Ethics & Global Politics*, Vol. 18, No. 3, 2025; Zhixiong Huang and Haokun Yu, "De Jure Necessity vs. de Facto Improbability? Establishing a New Treaty for Artificial Intelligence in the Military Domain", paper presented at the CIL/NUS AI Manual Workshop, Singapore, 2026.

[30] Charlotte Ku, "The Archipelagic States Concept and Regional Stability in Southeast Asia", *Case Western Reserve Journal of International Law*, Vol. 23, No. 3, 1991.

context-specific ethical structures. This integrated approach ensures the analysis of Regional Power dynamics in AI governance and the evolution of global norms.

5. AI Flaws and Their Impact on IHL Compliance

Let us consider a hypothetical scenario inspired by real-world events, such as the South China Sea conflict, to illustrate the possibility of AI flaws that may have a substantial impact on the civilian population. An AI-enabled autonomous drone swarm violates the distinction principle. It kills civilians when it incorrectly perceives fishing vessels as hostile because of an algorithmic error or misalignment in vessel recognition. Furthermore, opacity apparently hindered complete accountability, exposing value misalignment, as evidenced by confirmed US drone attacks in the Middle East, when AI-assisted targeting caused disproportionate injury despite claims of precision.[31] Another example is AI decision assistance in asymmetric urban combat, where sycophantic compliance with unlawful human orders leads to precautionary failures or even the endorsement of misaligned and extreme targeting suggestions that could instigate humanitarian catastrophes, including mass civilian casualties and indiscriminate attacks on innocent populations, without anyone being held accountable.[32]

In this hypothetical South China Sea scenario, the conflict will likely be asymmetric, taking the form of a grey zone conflict, which is neither peaceful relations nor armed conflict and occurs below the threshold of armed conflict.[33] In this context, the use of military AI could expose AI's vulnerabilities, as it may favour technologically superior forces, leading to unexpected escalation and violations of IHL's precautionary framework. However, less powerful forces may exploit these vulnerabilities to their advantage, risking violations of IHL. These erosions are especially noticeable in this hypothetical asymmetric grey zone conflict, where both sides amplify the risks and limitations of their military AI systems. The more technologically superior side often deploys AI to exploit its advantages but struggles with complex challenges, such as data poisoning, hacking, or poor sensor performance in cluttered environments.[34] Meanwhile, the weaker side uses tactics to reduce AI's claimed targeting accuracy by blurring distinctions and deliberately mixing with civilians or locating in the vicinity of civilian areas.[35] These hypothetical

---

[31] Kristina Benson, "'Kill 'em and Sort It out Later:' Signature Drone Strikes and International Humanitarian Law", *Global Business & Development Law Journal*, Vol. 27, No. 1, 2014, p. 36; Human Rights Watch, above note 10.

[32] Jonathan Kwik, "Digital Yes-Men: How to Deal with Sycophantic Military AI?", *Global Policy*, Vol. 16, No. 3, 2025.

[33] Javier Jordan, "International Competition Below the Threshold of War: Toward a Theory of Gray Zone Conflict", *Journal of Strategic Security*, Vol. 14, No. 1, 2021.

[34] Viacheslav Biletskyi, Viktor Tyshchuk and Oleksandr Mandziuk, "Autonomous Systems and the Speed of Battle: Legal Risks and Strategic Adaptation in AI-Enabled Warfare", *Scandinavian Journal of Military Studies*, Vol. 9, No. 1, 2026.

[35] ICRC, *Artificial Intelligence and Machine Learning in Armed Conflict: A Human-Centred Approach*, International Review of the Red Cross, Cambridge University Press, 2020.

incidents underscore the urgent need for regulatory action to bring AI into compliance with IHL and to stop systematic disruptions in armed engagements.

Despite the benefits it offers, applying artificial intelligence (AI) to military operations poses substantial challenges that might compromise compliance with IHL. The design, training, and operational deployment of AI systems in warfare are hampered by several fundamental problems. Moreover, when societal biases are reflected in training data, they can lead to biased decision-making, such as incorrectly identifying individuals across different populations based on their gender, race, or ethnicity, further amplifying the problems.[36] For example, non-Western populations have shown higher error rates in facial recognition algorithms, increasing the risks in multicultural conflict zones.[37] These vulnerabilities stem from development methods dominated by small datasets and profit-driven agendas, making them both technical and systemic problems.

To elaborate on the fundamental problems mentioned previously, this paper identifies four inherent flaws in AI systems in the military domain, particularly regarding LAWS and AI-DSS. These flaws are widely discussed as fundamental barriers to safe, ethical, and accountable deployment of AI-based weapons and support systems.[38] The first is the emergence of algorithmic biases that are sourced from skewed data training, flawed assumptions of the built machine learning model, which, in turn, can result in systematic misinterpretations that draw the line in discriminating legitimate from illegitimate targets, such as civilians or protected objects from combatants and military targets during armed conflict and inaccurate targeting, such as identifying neutral actors as threats due to poor pattern recognition in crowded regions, let alone with gender and racial biases.[39]

These problems may intensify in contexts of highly diverse ethnicities and cultures, and dense populations living in the vicinity of the military installation, such as those in Southeast Asia. The region, encompassing countries like Indonesia, the Philippines, Malaysia, Myanmar, and Vietnam, is home to hundreds of ethnic groups, languages, religions, and distinct social practices, creating a uniquely complex operational environment for military AI systems, particularly when the AI system is developed in a small number of countries with datasets training based on homogeneous populations, often those in Western and East Asian centric demographics.[40] This power imbalance risks solidifying biases in AI systems and possibly exporting defective technologies that compromise IHL in recipient States.

---

[36] L. Bruun and M. Bo, above note 11; J. Buolamwini and T. Gebru, above note 11.
[37] Ido Rosenzweig and Magdalena Pacholska, "The Use of Facial Recognition for Targeting under International Law", *International Review of the Red Cross*, Vol. 107, No. 928, 2025.
[38] Ingvild Bode, "Falling under the Radar: The Problem of Algorithmic Bias and Military Applications of AI", *Humanitarian Law & Policy*, 14 March 2024, available at: https://blogs.icrc.org/app/uploads/sites/102/2024/03/falling-under-the-radar-the-problem-of-algorithmic-bias-and-military-applications-of-ai-PDF.pdf.
[39] I. Bode, "Falling under the Radar" above note 38; L. Bruun and M. Bo, above note 11.
[40] I. Bode, "Falling under the Radar" above note 38.

Moreover, the already biased systems have the potential to misclassify individuals when data relies solely on appearance, behaviour, or geospatial patterns, particularly when the input data is incomplete or skewed, thereby undermining the distinction principle of IHL.[41] Unlike occasional technical failures, discriminatory algorithms are systemic and take much longer to fix. Once deployed, biased systems can cause widespread, repeated harm to vulnerable populations, such as ethnic minorities and densely populated communities.[42]

Second, the so-called "black boxes" are commonly characterized as AI systems that limit transparency and explainability of how algorithms reach decisions. Regarding LAWS and other autonomous systems, these systems must adhere to the fundamental principles of predictability (what the system will do) and understandability (why it does it), which are essential for the safe, legal, and responsible use of military AI under IHL.[43] Nevertheless, AI systems are inherently unpredictable due to brittleness (failure to perform on edge cases or out-of-distribution data), complex environments, adversarial inputs, self-learning capabilities, and interactions in swarms, which in turn create an "opaque" black box.

In the context of targeting decisions, the opacity of black boxes adds to the complexity of efforts to legally review any system under Article 36 of Additional Protocol I to the Geneva Conventions, which requires States to determine whether new weapons comply with international law.[44] Moreover, it is challenging to audit or challenge judgments in real-time combat scenarios due to the black-box opacity of AI models, which lack transparency and allow sophisticated neural networks to generate outputs without explicable logic.[45] When it is important for a field commander to understand the reasons and implications behind any decision it makes during the battle, the lack of ability to fully comprehend how an algorithm reached its final decision in targeting and engaging would severely compromise the meaningful role of human assessment – and thus the precautionary obligations to conform to international law.

Third, a significant reduction of human control and agency in increasingly autonomous systems jeopardizes accountability. The original foundation of IHL rests on the role of human agency, which places responsibility for any unlawful attack directly on the individuals and States responsible for its conduct. Nevertheless, when lethal decisions are delegated to algorithmic processes, identifying legal

---

[41] L. Bruun and M. Bo, above note 11.
[42] Ishmael Bhila, "Putting Algorithmic Bias on Top of the Agenda in the Discussions on Autonomous Weapons Systems", *Digital War*, Vol. 5, No. 3, 2024.
[43] Arthur Holland Michel, "The Black Box, Unlocked: Predictability and Understandability in Military AI", *UNIDIR*, Geneva, 2020.
[44] Protocol Additional (I) to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977 (entered into force 7 December 1978), Art. 36.
[45] J. Kwik and T. Van Engers, above note 12.

responsibility becomes complex, eliciting concerns of accountability gaps and potential impunity.[46] Such gaps threaten the enforceability of core principles of distinction, proportionality, and precaution under the Geneva Conventions and their Additional Protocols.[47]

Furthermore, empirical research has shown that AI-based systems can increase the likelihood of non-compliant behavior in simulated scenarios when human involvement is absent.[48] Moreover, legal experts contend that AI operations with less human oversight obscure accountability and encourage impunity for IHL violations[49] and that accountability deteriorates in the absence of required human-in-the-loop procedures, potentially leading to a division of accountability among States, developers, and operators.[50] Henceforth, tracing accountability becomes difficult when choices are left to autonomous systems, as operators may rely on AI results unquestioningly, resulting in an "accountability gap."[51]

The fourth and final is the emergence of value misalignment between the AI's learning and optimization processes, which diverge from humanitarian ethical values and norms, including IHL, thereby deepening ethical gaps and tensions. The emergent misalignment could result from the black-box nature of how the machine learns from its data and reaches decisions.[52] Sycophantic AI is essentially a behavioral symptom of an AI system's misaligned objective, which often amplifies extreme or misaligned commander inputs instead of correcting them.[53]

In some cases, AI systems may ignore alternate strategies that promote humanitarian outcomes and compromise precautionary obligations or even suggest an extreme measure to maximize operational gains and overall mission accomplishment metrics, rather than to consider and assess humanitarian or ethical

---

[46] Will Fleisher *et al.*, "Responsibility and Accountability in an Algorithmic Society", *Philosophy & Technology*, Vol. 38, No .4, 2025; I Nyoman Agus Prabawa *et al.*, "Legal and Ethical Implications of Algorithmic Decision-Making Systems: A Critical Analysis of Challenges and Regulatory Solutions", *International Journal of Law*, Vol. 11, No. 8, 2025, p. 6.

[47] Jean-Marie Henckaerts, Louise Doswald-Beck and Carolin Alvermann (eds), *Customary International Humanitarian Law, Vol. 1: Rules*, Cambridge University Press, Cambridge, 2009; E. Winter, above note 12.

[48] Ujué Agudo *et al.*, "The Impact of AI Errors in a Human-in-the-Loop Process", *Cognitive Research: Principles and Implications*, Vol. 9, 2024; Toby Drinkall, "Red Lines and Grey Zones in the Fog of War: Benchmarking Legal Risk, Moral Harm, and Regional Bias in Large Language Model Military Decision-Making", Preprint, arXiv, 2025, available at: https://doi.org/10.48550/ARXIV.2510.03514.

[49] Nanda Min Htin and Shriya Srikanth, "Redefining the Standard of Human Oversight for AI Negligence", *Harvard Journal of Law & Technology*, 9 February 2026.

[50] Ilse Verdiesen, Filippo Santoni de Sio and Virginia Dignum, "Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight", *Minds and Machines*, Vol. 31, No. 1, 2021.

[51] Rebecca Crootof, "AI and the Actual IHL Accountability Gap", *Centre for International Governance Innovation*, 28 November 2022, available at: https://www.cigionline.org/ articles/ai-and-the-actual-ihl-accountability-gap/.

[52] R. Lindelauf and H. Meerveld, above note 4.

[53] J. Kwik, "Digital Yes-Men" above note 32.

standards, even if that would mean committing a grave violation of existing rules and norms.[54] For example, the misalignment can manifest in suggestions to prioritize military efficiency and operational advantage over moral accountability and IHL compliance during *in bello* situations, and even, unethically, in the spoofing of IHL-protected entities, such as ICRC-affiliated systems.[55]Hence, the problem lies beyond a mere technical malfunction. It stems from a structural discrepancy between machine optimization and human moral standards and ethical reasoning.[56]

When these flaws take place, they directly undermine the fundamental tenets of *jus in bello*, embodied in IHL, which obligate humane conduct during the time of conflict. A clear distinction between combatants and civilians is necessary to uphold the principle of distinction. Furthermore, the opacity of the black box may compromise proportionality, which balances anticipated human suffering against projected military advantage, allowing AI to execute excessive strikes without verifiable assessments of collateral damage. Meanwhile, IHL expects commanders on the field to take reasonable steps to eliminate uncertainty, likely stemming from the opacity of black boxes, by mitigating civilian collateral damage.[57]

Furthermore, the emergence of an opacity characteristic potentially widens this gap, making it more difficult to conduct post-incident investigations and to bring charges under IHL procedures, such as those of the International Criminal Court. Even further, Biased AI may allow for recurrent misidentifications without corrective feedback loops, leading to systematic violations and the continuance of humanitarian crises. Impunity due to the obscurity of AI accountability risks normalizing AI-driven excesses in areas with weak enforcement, such as several regions of the Indo-Pacific, which further undermines trust in international rules. This precedent is inferred from the regional historical proclivity toward structural, institutional, geopolitical, and power-based factors that limit the application and enforcement of international rules, particularly in emerging domains such as AI or military technology governance in relation to the implementation of IHL.[58] Regulations and legal frameworks are already lagging behind the emergence of new military technologies, let alone the weak enforcement, which will only further complicate

---

[54] L. Bruun and M. Bo, above note 11; ICRC, "Investigating and Prosecuting Serious Violations" above note 17; Juan-Pablo Rivera *et al.*, "Escalation Risks from Language Models in Military and Diplomatic Decision-Making", *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, 2024; Taylor Kate Woodcock, "Human/Machine(-Learning) Interactions, Human Agency and the International Humanitarian Law Proportionality Standard", *Global Society*, Vol. 38, No. 1, 2024.
[55] J. Guo, above note 29.
[56] Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Lieber Institute West Point*, available at: https://lieber.westpoint.edu/can-ai-teach-itself-abuse-ihl-protected-indicators/.", *Lieber Institute West Point*, 5 December 2025.
[57] Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies*, TMC Asser Press, The Hague, 2024.
[58] Grace Jeeann Ha, "Weak Enforcement, Flawed Procedure, Dangerous Precedent: Analyzing the Effectiveness of International Law Through the Lens of the South China Sea Dispute", *The Undergraduate Law Review at UC San Diego*, Vol. 3, No. 1, 2025; Rithiya Serey, "ASEAN's Ineffective Response to The South China Sea Disputes", Thesis, Northern Illinois University, 2021.

their implementation. This situation is often critiqued as the "paper tiger" when norms and adjudication are strong but weak on execution and enforcement due to normative factors, as mentioned above.[59]

6. Global and Regional Military AI Governance: Vulnerabilities and Opportunities

Due to conflicting interests and rapidly advancing technologies, the global governance of military AI applications remains fragmented. The current development of military AI is heavily influenced by Major Powers, particularly the US and China, which often prioritize national security and competitive advantage over broader moral or legal requirements. Adopting the AI principles outlined in its Ethical AI guidelines, the US Department of Defence grants flexibility while prioritizing responsible innovation to attain operational superiority. This innovation includes investments in autonomous systems, like Project Maven, for targeted optimization.[60] Similarly, to preserve an advantage in the Indo-Pacific, China's "civil-military fusion" policy incorporates AI into its People's Liberation Army, focusing on rapid deployment in areas such as surveillance and unmanned vehicles, often with less emphasis on transparency.[61]

Meanwhile, this trend is also evident in Russia's AI advancements in electronic warfare and hypersonic systems, which force Middle and Small Powers to either bandwagon—adopting technologies from these leaders—or pursue expensive, independent development paths.[62] Realistically speaking, this extended dominance from the Superpowers creates a global security conundrum in which AI arms races undermine collective efforts to enact humanitarian protections. With that context as the strategic backdrop, this paper argues that multilateral measures are essential to govern military AI effectively and to promote ethical application and adherence to IHL. Since 2017, the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS) has been convened by the Convention on Certain Conventional Weapons (CCW) under the auspices of the United Nations. The GGE's mission is to promote protocols that guarantee meaningful human

---

[59] Jonathan Alberts, "Paper Tiger of the Seas: The South China Sea and the Enforcement Gap in UNCLOS", *American University International Law Review*, 21 January 2026, available at: https://auilr.org/2026/01/21/paper-tiger-of-the-seas-the-south-china-sea-and-the-enforcement-gap-in-unclos/.

[60] US Department of Defence, *DOD Adopts Ethical Principles for Artificial Intelligence* Washington, DC, 24 February 2020, available at: https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[61] Cole McFaul, Sam Bresnick and Daniel Chou, *Pulling Back the Curtain on China's Military-Civil Fusion*, CSET, Washington, DC, 2025.

[62] Kateryna Bondar, *How Russia Is Reshaping Command and Control for AI-Enabled Warfare*, Centre for Strategic and International Studies, Washington, DC, 2026, available at: https://www.csis.org/analysis/how-russia-reshaping-command-and-control-ai-enable d-warfare.

control, transparency, and legal reviews of AI systems.[63] Some of the guiding principles suggested in these discussions include requiring bias audits to reduce the risk of discrimination and prohibiting fully autonomous weapons that do not adhere to the IHL principles of differentiation and proportionality.

Further, to align AI with IHL's precautionary obligations, the Netherlands and South Korea launched the Responsible AI in the Military Domain (REAIM) initiative in 2023. This initiative emphasizes ethical design, human-centric controls, and international cooperation, and it encourages a multi-stakeholder approach.[64] Currently, 61 States have endorsed REAIM's summits, which focus on practical solutions, including standard best practices for AI deployment and testing in asymmetric environments.[65] Standardized norms, including interoperability standards to prevent and mitigate value misalignment, are supported by other initiatives, including NATO's AI Strategy and the EU's proposed AI Act, which includes military exemptions and ethical guidelines.[66] Going back to the theoretical framework, constructivist perspectives show how these platforms might socially generate norms and advance IHL by overcoming AI's bias and opacity through dialogue and consensus.

Despite these advancements, current governing frameworks face considerable limitations that obstruct effective IHL compliance.[67] Due to major nations' resistance to LAWS bans to maintain technological advantages, the CCW's consensus-based decision-making process has stalled the development of legally binding treaties, resulting in non-binding guiding principles rather than enforceable protocols.[68] REAIM, even though innovative, lacks universal ratification and enforcement procedures, relying on voluntary commitments that may not deter violations in intense conflicts.[69] Moreover, frameworks frequently ignore the

---

[63] 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, CCW/GGE.1/2025/WP.4, 21 August 2025, available at: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_%282025%29/CCW-GGE.1-2025-WP.4.pdf.

[64] Netherlands Ministry of Foreign Affairs, *AI in the Military Domain: The Netherlands Advocates for International Dialogue*, 25 September 2025, available at: https://openrijk.nl/en/ministeries/buitenlandse-zaken/artikel/0bbc80a4-3442-4bfc-8efe-d993eb2743e5/ai-in-the-military-domain:-the-netherlands-advocates-for-international-dialogue; Republic of Korea Ministry of Foreign Affairs, *Outcome of Responsible AI in Military Domain (REAIM) Summit 2024*, 9 October 2024, available at: https://overseas.mofa.go.kr/eng/brd/m_5676/view.do?seq=322676.

[65] Republic of Korea Ministry of Foreign Affairs, above note 63.

[66] European Commission, "AI Act", *Shaping Europe's Digital Future*, 20 February 2026, available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[67] A. Batool *et al.*, above note 13.

[68] Ray Acheson (ed.), *Disarmament Is Always the Smartest Choice*, CCW Report, Vol. 13, No. 2, Reaching Critical Will Programme, 2025.

[69] Denise Garcia, "The Global Diplomacy of Governing Military Artificial Intelligence", *Ethics and Information Technology*, Vol. 27, No. 4, 2025; H. Meerveld *et al.*, above note 13.

specificity of archipelagic vulnerabilities, such as AI's role in maritime surveillance. Where bias could instigate and even escalate conflicts without accountability, creating gaps in applicability to asymmetric warfare, which is prevalent in the Indo-Pacific. Additionally, the proliferation of black-box systems, which sustain impunity and misaligned values, is enabled by the lack of mandatory transparency requirements. Civil society criticisms, such as those from Human Rights Watch and the Campaign to Stop Killer Robots, contend that these restrictions permit the systematic undermining of IHL, particularly in areas with limited influence of international ethical standards.[70] To address them, academics support hybrid strategies that preserve inclusivity in norm-setting for Middle Powers by combining progressive binding measures with soft law.[71]

At the regional level, there is a unique combination of geopolitical tensions, maritime vulnerabilities, and opportunities for AI governance in the military domain across the Indo-Pacific, especially in Southeast Asia. Escalating maritime disputes are a defining feature of this region, particularly in the South China Sea (SCS), where heavily populated archipelagic States meet overlapping territorial claims from China, the Philippines, Vietnam, Malaysia, and Brunei.[72] These disputes involve not only resource-rich exclusive economic zones but also important sea lanes vital for global trade, with over $5 trillion in annual commerce traversing the area.[73] The archipelagic geography of Southeast Asia, with its thousands of islands and diverse inhabitants, increases vulnerabilities because conflicts often occur near civilian settlements, raising the risk of humanitarian disasters in asymmetrical settings. China has escalated tensions, shifted the strategic balance, and challenged the regional order through the construction of artificial islands and the deployment of cutting-edge surveillance technologies, including AI-enabled systems.[74] In this context, the use of AI in military operations—such as surveillance and unmanned ships—exacerbates gray zone tactics, in which non-lethal but coercive measures make it harder to distinguish between peace and violence. The scenario is further complicated by the area's ethnic and religious diversity, as misidentifications in densely populated areas could result in disproportionate harm to civilians.

---

[70] 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*, 13 January 2026.

[71] Philip Alexander, "Reconciling Automated Weapon Systems with Algorithmic Accountability: An International Proposal for AI Governance", *Harvard International Law Journal*, 16 October 2023, available at: https://journals.law.harvard.edu/ilj/2023/10/ reconciling-automated-weapon-systems-with-algorithmic-accountability-an-international-proposal-for-ai-governance/; Zeynep Orhan et al., "Artificial Intelligence Standards in Conflict: Local Challenges and Global Ambitions", *Standards*, Vol. 5, No. 4, 2025; Scott Sullivan, "Targeting in the Black Box: The Need to Reprioritize AI Explainability", *Lieber Institute West Point*, 28 August 2024, available at: https://lieber.westpoint.edu/targeting-black-box-need-reprioritize-ai-explainability/.

[72] A. Triki, above note 9.

[73] S. Chhaya, above note 9.

[74] Fumiko Sasaki, "Space, Maritime Security, and Geopolitics in the South China Sea", *Journal of Indo-Pacific Affairs*, 2025.

The use of AI in military contexts increases preexisting vulnerabilities, which could intensify conflicts and undermine Southeast Asia's adherence to IHL. In the SCS, AI-enabled systems, such as autonomous drones and water cannons, could lower the threshold for confrontation, enabling precise yet escalatory gray zone operations that disproportionately harm civilians across diverse settings.[75] Ethnic tensions could be exacerbated by misalignment hazards, such as biased algorithms in surveillance, leading to misidentifications and violations of the norms of proportionality and distinction.[76] Additionally, deepfakes and AI-driven propaganda amplify narratives for information warfare, skewing public opinion and possibly galvanizing support for military expansion in asymmetric conflicts. These consequences heighten humanitarian risks, underscoring the need for region-specific protections to prevent AI from instigating and escalating crises and perpetuating impunity.

In the midst of Great-Power tensions, ASEAN plays a crucial role in promoting context-specific AI governance by highlighting the importance of norm-building. Although it does not address military applications, the 2024 ASEAN Guide on AI Governance and Ethics establishes the framework for regional ethical frameworks that promote openness and diversity. Harmonized guidelines, dialogues with Major Powers such as the US and China, and non-lethal AI applications for maritime domain awareness are promoted by initiatives such as the Working Group on AI Governance and the ASEAN Defence Ministers' Meeting (ADMM) Joint Statement on AI Cooperation.[77] To ensure AI complies with cautious IHL requirements in asymmetric conflict, ASEAN may reconcile global norms with regional needs by leveraging platforms such as the East Asia Summit and the ASEAN Regional Forum (ARF).

As Middle- and Small-Power nations, Southeast Asian nations must balance the risks and opportunities of bandwagoning the AI Superpowers, China and the United States, or pursuing independent paths.[78] While indigenous development promotes sovereignty but suffers from resource shortages, bandwagoning provides access to cutting-edge technologies but risks dependency and value misalignment. These problems could be resolved with a well-rounded strategy that includes

---

[75] James Black *et al.*, *Strategic Competition in the Age of AI: Emerging Risks and Opportunities from Military Use of Artificial Intelligence*, RAND Corporation, Santa Monica, CA, 2024; Ashish Dangwal, "China Readies World's 1st AI-Enabled Water Cannon That It Claims Can Revolutionize Non-Lethal Combat", *Eurasian Times*, 13 April 2024, available at: https://www.eurasiantimes.com/china-readies-worlds-1st-ai-enabled-water-canon/.

[76] L. Bruun and M. Bo, above note 11.

[77] Muhammad Faizal Bin Abdul Rahman, "AI in Defence: The Way Forward for ASEAN's ADMM Cooperation", RSIS, August 2025; ASEAN Secretariat, *Expanded ASEAN Guide on AI Governance and Ethics Generative AI*, 2024; ASEAN Secretariat, *Joint Statement by the ASEAN Defence Ministers on Cooperation in the Field of Artificial Intelligence in the Defence Sector*, 26 February 2025; B.A. Putra, above note 14.

[78] Wen Zha, "Southeast Asia amid Sino-US Competition: Power Shift and Regional Order Transition", *The Chinese Journal of International Politics*, Vol. 16, No. 2, 2023.

minilateral alliances, AI supply chain specialization, and hedging through diversification, enhancing regional resilience and IHL compliance.

7. Discussion and Analysis

The application of AI in warfare compromises ethical judgment and accountability, undermining the core principles of IHL. By mistakenly categorizing civilians as combatants, algorithmic bias—which often originates from skewed training data that embodies societal prejudices—leads to discriminatory targeting, especially in diverse communities, and violates the principle of distinction. Explainability is hampered by black-box opacity, which makes it hard to validate proportionality evaluations that require expected military benefits to exceed civilian loss. This results in disproportionate strikes that lack traceable justification. While value misalignment departs from human ethical standards, aggravating precautionary failures and maintaining a responsibility gap that sustains impunity for transgressions, sycophantic inclinations magnify defective inputs.[79] These shortcomings risk systematic violations of IHL in the absence of legally binding standards, exacerbating humanitarian crises in both symmetric and asymmetric conflicts. While realism emphasizes the Major Powers' aversion to restrictions that would limit their strategic advantages and so prolong governance gaps, constructivism emphasizes how norms are built and changed through multilateral discourse. This association underscores the essential need for inclusive frameworks that align AI with humanitarian needs.

In Southeast Asia's asymmetric conflicts, marked by maritime disputes and archipelagic vulnerabilities, AI-driven deficiencies intensify power imbalances and humanitarian issues. Grey zone tactics in the South China Sea, where biased algorithms misidentify civilian vessels, could intensify as AI mismatches in surveillance and targeting deepen. These vulnerabilities would violate the distinction and proportionality requirements in highly populated, ethnically diverse areas. Reduced human control widens accountability gaps, allowing non-state actors to operate with impunity in hybrid warfare and intensifying problems amid cultural diversity. As Middle Powers, such as Indonesia, acquire potentially biased AI from Superpower countries, they are increasingly heightening disparities and, at the same time, undermining regional order without the presence of ASEAN-centric ethical norms. In the absence of such norms, realist dynamics emerge as they manage bandwagoning challenges toward the Superpowers. According to constructivism, ASEAN's prominence could encourage context-specific norms, but long-standing power imbalances risk sustaining infractions in asymmetrical maritime contexts. This situation highlights the heightened challenges in the Indo-Pacific, where, in the absence of specific regulation for the emerging technologies, AI could cause excessive harm to vulnerable populations.

[79] J. Kwik, "Digital Yes-Men" above note 32.

While underlining adaptable IHL issues in Southeast Asian maritime contexts, insights from the deployment of AI drones in Middle Eastern wars, such as those in Gaza and Yemen, reveal universal hazards. Due to bias and opacity, AI-assisted targeting in the Middle East has disproportionately harmed civilians,[80] which is comparable to possible escalation in Indo-Pacific conflicts, where surveillance errors could mistakenly recognize a fishing vessel as a warship amid territorial claims.[81] In asymmetric warfare, both regions struggle with accountability gaps; however, unlike the Middle East's urban-centric operations, Southeast Asia's archipelagic diversity and predominant maritime background may heighten the risk of bias and hazards. Swarms of autonomous drones in the Middle East serve as a reminder of the growing dangers associated with autonomous systems. However, Indo-Pacific vulnerabilities highlight maritime-specific issues, requiring contextually adaptable governance to prevent the erosion of IHL, which is generally prevalent in the region.

Proponents of AI implementation during war contend that by outperforming human error in targeting and decision-making, it increases precision and lowers collateral damage. This argument is refuted by empirical evidence showing that biases in training data erode precision across varied or asymmetric environments and sustain discriminatory effects. Large Language Models (LLMs) acting as military decision-agents in the AI-DSS framework have potential behavioral risks that produce IHL-violating outputs—including civilian targeting—in 16.7% to 66.7% of simulated scenarios when human oversight is reduced or absent.[82] Similar evidence also emerged regarding errors related to ethnic or racial recognition. The NIST Face Recognition Vendor Test (FRVT) tests 189 algorithms from 99 developers on massive datasets, showing that most algorithms exhibit demographic differences: false-positive rates for African American and Asian faces are 10 to 100 times higher than for Caucasian faces in one-to-one matching. Moreover, Algorithms trained predominantly on Western/European datasets performed worse on non-Western populations, with performance degrading sharply under asymmetric conditions, including differences in lighting, poses, or ethnic distributions.[83]

Additionally, one study shows that error rates were particularly severe for darker-skinned females (up to 34.7% vs. 0.8% for lighter-skinned males in an earlier study of gender shades benchmarks).[84] Another study also finds significantly higher word error rates for Black speakers than for White speakers, even after controlling for other variables. Similar findings appear in studies on non-

[80] K. Benson, above note 31; Human Rights Watch, above note 10.
[81] James Johnson, *Artificial Intelligence and the Future of Warfare: The USA, China, and Strategic Stability*, Manchester University Press, Manchester, 2021.
[82] T. Drinkall, above note 48.
[83] Patrick Grother, Mei Ngan and Kayee Hanaoka, *Face Recognition Vendor Test Part 3: Demographic Effects*, NIST IR 8280, National Institute of Standards and Technology, Gaithersburg, MD, 2019, available at: https://nvlpubs.nist.gov/nistpubs/ir/2019/ NIST.IR.8280.pdf.
[84] J. Buolamwini and T. Gebru, above note 11.

Western accents and dialects. These biases persist and amplify in varied acoustic or linguistic environments outside the training distribution.[85]

Furthermore, automation bias increases reliance on flawed systems, raising rather than lowering risks, as seen in real-world deployments where opacity led to misidentifications. Extensive testing reveals persistent defects, increased error rates, and non-compliant behavior, proving that AI's "rational" efficacy falters without human supervision. For example, simulation-based experiments with large language models (LLMs) as agents in conflict scenarios demonstrate IHL violation rates (e.g., civilian targeting) ranging from 16.7% to 66.7% when human involvement is absent or limited. Automation bias further compounds this: operators tend to over-rely on AI recommendations, reducing critical intervention even when defects are present.[86] Moreover, autonomous systems remain "highly prone to error," with poor robustness, interpretability, and vulnerability to adversarial attacks. Extensive testing in realistic operational environments (including countermeasures) reveals failures in perception, generalization, and decision-making that persist despite iterative improvements. Systems often overfit to narrow test scenarios and falter under distribution shifts common in asymmetric warfare.[87] These empirical findings highlight the need for legally mandated standards to disprove inflated claims of accuracy.

8. Recommendations and Policy Implications

Based on the discussion and analysis above, this paper proposes several recommendations that are expected to impact policy-making. They could shape the implementation of AI-enabled weapon systems during open conflict or even war. These recommendations focus on realistic, achievable steps at the multilateral, regional, and operational levels to realign AI with ethical standards, especially in light of Southeast Asia's maritime vulnerabilities. They build on the paper's main arguments, which are that AI flaws like bias, opacity, and misalignment undermine IHL's human-centric jus in bello principles (distinction, proportionality, precaution), promote impunity, and amplify humanitarian crises in asymmetric conflicts without binding norms.

First, to encourage multilateral action on leveraging binding changes to the CCW protocols that call for "human-in-the-loop" controls on AI targeting to guarantee IHL compliance and reduce opacity-induced violations. This effort can be

[85] I. Bhila, above note 42.
[86] Vanessa Vohs, "Meaningful Human Agency in Automated Weapon Systems: A Plea for Human-in-the-Loop Regulation", *LSE Law Review Blog*, 9 October 2021, available at: https://blog.lselawreview.com/2021/10/09/meaningful-human-agency-in-automated-weapon-systems-a-plea-for-human-in-the-loop-regulation/.
[87] A. Holland Michel, above note 43; Shayne Longpre, Marcus Storm and Rishi Shah, "Lethal Autonomous Weapons Systems & Artificial Intelligence: Trends, Challenges, and Policies", *MIT Science Policy Review*, Vol. 3, 2022.

proposed and done through UN GGE sessions. This recommendation proposes a two-track multilateral strategy: (1) pushing for binding amendments to CCW protocols mandating "human-in-the-loop" or "meaningful human control" on AI targeting systems to ensure IHL compliance and reduce opacity-related violations, advanced through UN GGE sessions; and (2) accelerating acceptance of the Responsible AI in the Military Domain (REAIM) initiative to address accountability gaps and value misalignment, supported by inclusive ratification campaigns to counter Great-Power resistance. Nevertheless, only three ASEAN member countries – Malaysia, the Philippines, and Singapore – are participating members of the UN GGE regarding LAWS.[88] Meanwhile, Thailand has participated as an observer and joined joint statements in 2025. Indonesia, Vietnam, and Cambodia show limited or no consistent engagement in GGE sessions, although some support for a broader ASEAN statement in the UNGA First Committee.[89]

In relation to REAIM, it is crucial to expedite its acceptance through diplomatic endorsements to address accountability gaps and misalignment of values. REAIM emphasizes alignment with human values, IHL, and ethical standards throughout the AI lifecycle. It encourages bias audits, explainability, and testing in diverse contexts — countering sycophantic or culturally insensitive AI behaviours.[90] Moreover, it stresses human responsibility, traceability, and oversight mechanisms, helping diffuse responsibility issues by clarifying roles for developers, operators, and commanders.[91] Endorsing REAIM provides a practical framework that can inform CCW negotiations, building momentum for stronger norms without the immediate veto authority of the Great-Powers. For that reason, international diplomatic and civil society must aim to broaden the endorsement and inclusive ratification of the REAIM Blueprint for Action (BFA).[92]

This first recommendation may be low on feasibility in the short term. However, it is practically feasible as a complement to the existing norms and viable as a long-term, incremental norm-building strategy. The low feasibility is due to the challenge of securing binding CCW changes that would face significant resistance.

---

[88] Manoj Harjani, "The REAIM 'Blueprint for Action' Needs Skin in the Game", *IDSS Paper*, Singapore, 16 January 2025; Ousman Noor, "70 States Deliver Joint Statement on Autonomous Weapons Systems at UN General Assembly", *Stop Killer Robots*, 21 October 2022, available at: https://www.stopkillerrobots.org/news/70-states-deliver-joint-stateme nt-on-autonomous-weapons-systems-at-un-general-assembly/.

[89] Convention on Certain Conventional Weapons - Group of Governmental Experts on Lethal Autonomous Weapons Systems (2026), United Nations Office for Disarmament Affairs, 2026, available at: https://meetings.unoda.org/ccw-/convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2026.

[90] Yasmin Afina, *The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain*, Summary Report, UNIDIR, Geneva, 2026.

[91] 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, The Hague, September 2025.

[92] M. Harjani, above note 88.

Nevertheless, combining targeted GGE proposals with REAIM diplomatic endorsements and inclusive coalitions offers a pragmatic path forward that is perceived as a middle way and better options for the majority of countries, particularly the global south, which consists of Middle- and Small-powers. Henceforth, ASEAN should prioritize capacity-building to expand GGE participation and safeguard regional interests.

Second, to prevent an escalation in the South China Sea and to protect diverse civilian groups in one of the world's most ethnically and culturally heterogeneous regions, ASEAN should proactively develop tailored ethical principles for military AI. These principles must explicitly address the unique risks of algorithmic bias in maritime surveillance and targeting systems operating across archipelagic environments, dense coastal populations, and contested waters. Specifically, bias audits for maritime surveillance must be incorporated, covering training data representativeness (e.g., behavioural patterns, traditional maritime activities), performance testing in varied Southeast Asian contexts, and regular third-party evaluations. Concurrently, the existing ASEAN Guide on AI Governance and Ethics (2024) — currently focused on civilian and dual-use applications — must be expanded through inclusive consensus negotiations to include a dedicated module on responsible military AI use, incorporating requirements for meaningful human control, transparency in high-risk systems, and safeguards against value misalignment.

Additionally, building resilience against power asymmetries through established ASEAN-led platforms, such as the ARF for confidence-building dialogues (ARF for track 1.5/2 dialogues on confidence-building measures, simulations, and bias mitigation best practices), the ADMM and ADMM-Plus for practical cooperation rather than confrontation, is essential. This recommendation includes encouraging multilateral coalitions for cooperative and joint AI training, shared simulation exercises, and the development of shared datasets across Southeast Asian States to reduce over-reliance on external (the Great-Powers) AI systems and foster regional interoperability under common ethical standards. Essentially, the substantive idea aligns with this paper's primary theory of constructivism.

This second recommendation is realistic, though challenging, because it builds on ASEAN's institutional strengths, including foundational norms of consensus, non-interference, and incrementalism (the "ASEAN way"), which make binding treaties difficult but make voluntary or soft-law instruments highly achievable. The Declaration on the Conduct of Parties in the South China Sea (2002) and ongoing Code of Conduct (COC) negotiations demonstrate that ASEAN excels at producing joint statements, guidelines, and frameworks rather than hard treaties when facing Great-Power sensitivities.

Moreover, it also builds on recent momentum and existing frameworks as leverage points. The recent Joint Statement by ASEAN Defence Ministers on

Cooperation in the Field of AI in the Defence Sector (February 2025) already signals a willingness to engage on military AI, providing a ready foundation. A Joint Statement or ADMM Declaration on Responsible Military AI Use is the most feasible next step. Further, it can be elevated to a dedicated ASEAN Defence Ministers' Guideline or Protocol on Military AI (non-binding initially) if momentum builds.

Likewise, the 2024 ASEAN AI Governance Guide already establishes seven core principles (transparency, fairness, security, accountability, human-centricity, data governance, and reliability). Expanding it to military applications, through the ASEAN Digital Ministers' Meeting (ADGMIN) and then integrating military aspects via ADMM coordination, should be a low-resistance evolutionary step compared to creating new mechanisms from scratch. Ultimately, this recommendation positions ASEAN as a responsible regional actor capable of shaping military AI norms amid Great-Power rivalry, ultimately reducing escalation risks in the South China Sea while safeguarding civilian populations.

Third, ethical design standards, including mandatory pre-deployment bias audits, requirements for explainable AI (XAI) models trained on sufficiently large and representative datasets obligations, and robust testing protocols to mitigate both automation bias (user-side) and model sycophancy (system side) that potentially amplify misaligned commander inputs, particularly in multicultural and highly diverse operating environments such as Southeast Asia – should be enforced through State-led legal reviews and independent technical accreditation bodies, with the ICRC contributing the substantive humanitarian benchmarks and convening expert dialogue. This division preserves the ICRC's neutrality and matches its mandate. Such standards would help reduce the risk of IHL violations stemming from opacity, poor generalization in asymmetric maritime or urban settings, and cultural misinterpretations common in the South China Sea and broader Indo-Pacific region.

Simultaneously, promoting Middle- and Small-Power and ASEAN capacity-building should focus on IHL-training compliance training, legal-review methodology, and doctrine support through carefully managed open technology transfers and collaborative research initiatives, rather than on the transfer of the underlying technology, given export-control constraints and the acute sensitivity of operationally derived training data. These efforts would enable regional States to develop or adapt AI systems aligned with local contexts, reducing over-reliance on Great-Power technologies while fostering IHL-compliant innovation. These transfers must include strict safeguards for sensitive data, intellectual property protections, and mutual verification mechanisms to prevent misuse or proliferation.

This recommendation is moderately feasible and pragmatic as a complementary soft-law and capacity-building approach for Middle- and Small-Power countries, leveraging the ICRC's credibility and expertise, including its unparalleled legitimacy in IHL matters and extensive experience in assessing new

technologies through an "effects-based" approach. Furthermore, regionally tailored standards audits, guided by ICRC's recommendations and benchmarks, offer a neutral, humanitarian-focused pathway. Additionally, pre-deployment audits and XAI requirements directly address bias and sycophancy in multicultural contexts, as recommended by the ICRC's emphasis on testing, validation, bias mitigation, and preservation of human judgment. This suggestion arguably aligns with the ICRC's mandate and also with ASEAN's preference for consensus-based, non-confrontational mechanisms and counter asymmetries. Related to the contribution of this paper toward knowledge, the potential to conduct empirical research on AI simulations in asymmetric settings in Southeast Asia to assess the effects of bias on IHL and inform policy modifications holds promise for future research directions.

Furthermore, it is essential to investigate the legal effectiveness of new treaties in addressing opacity and to fill gaps in the literature through comparative regional studies. On top of that, the inclusivity of research can be enhanced by investigating non-Western ethical viewpoints in AI governance. The inclusion of diverse perspectives could enhance future research. Finally, these paths ensure practical advancements in AI-IHL alignment by supporting continuous normative evolution.

9. Conclusion

This paper has argued that the inherent structural vulnerabilities of artificial intelligence — algorithmic bias, black-box opacity, the erosion of meaningful human control, and value misalignment — systematically undermine the human-centric foundations of *jus in bello*. These are not incidental technical deficiencies that can be corrected at the margins of deployment; they are constitutive features of how current AI systems learn, generalize, and optimize. When embedded in lethal autonomous weapons systems and AI-enabled decision-support architectures, they erode the three principles upon which IHL's protection of civilians rests: distinction, proportionality, and precaution. In the absence of legally binding governance standards, these vulnerabilities do not describe theoretical risks — they describe operational realities in formation.

The constructivist analysis presented in this paper demonstrates that the international community is not passive in the face of these realities. The CCW Group of Governmental Experts, the REAIM Blueprint for Action, and the 2025 ADMM Joint Statement on AI Cooperation all reflect active norm-building. Nevertheless, the realist analysis demonstrates with equal clarity why that norm-building remains structurally insufficient: Major Military powers resist binding constraints that erode their competitive advantage, and the consensus architecture of multilateral forums converts normative aspirations into non-binding principles. Southeast Asia experiences this tension more intensely than other regions. Moreover, the region's archipelagic geography, ethnic and cultural heterogeneity, contested maritime domain, and Middle-Power strategic position place it at the precise intersection of

AI's greatest operational risks and international governance's deepest structural failures.

Quo vadis, then? Based on the recommendations proposed in this paper, Human-aligned IHL evolution means three concrete things that existing frameworks gesture toward but have not yet achieved. First, binding protocols — not voluntary principles — under the CCW establish mandatory meaningful human control as a threshold condition for autonomous targeting systems. Second, regional institutions, such as ASEAN, should proactively develop ethical principles for military AI, including the legal codification of developer and state co-responsibility for AI-driven IHL violations, and the elimination of the accountability gap that disperses legal responsibility among multiple actors, leaving no individual actor accountable. Likewise, the explicit prohibition of autonomous weapons systems that are inherently incapable of IHL compliance — systems that, by technical design, cannot apply the proportionality assessment or exercise the precautionary judgment mandated by IHL for any entity engaged in armed conflict. Third, mandatory pre-deployment IHL compliance audits by independent technical bodies guided by ICRC humanitarian benchmarks, assessing AI systems for bias, opacity, and value misalignment before operational deployment. Together, these elements do not eliminate AI from the battlefield. They constitute the legal architecture of a battlefield in which AI remains a tool of human decision-making rather than its surrogate.

For Southeast Asia, the projection of achieving successful regional norm-building is distinctly tangible. It would mean that the norm is to close the left gap in the defence domain, harmonize bias-audit standards, and reduce the risk of escalation that asymmetric AI errors might trigger in the maritime domain, particularly in the South China Sea. Furthermore, ASEAN member countries are expected to participate more as norm-contributors rather than norm-recipients within the CCW and REAIM processes, advancing the archipelagic and maritime-specific dimensions of IHL that land-based, Western-centric frameworks have consistently neglected. None of these outcomes requires the consent of the Great Powers. They would only require political will among States that already possess the institutional infrastructure — ASEAN, the ARF, the ADMM+ — and the strategic incentive to deploy it.

Fundamentally, this paper never meant "*Quo Vadis*" as a question about technology but more as a question about the legal perspective — whether the institutions humanity has built to limit the barbarism of war can adapt faster than the systems designed to wage it. IHL was built on the premise that a moral agent exists at the threshold of lethal force: an agent capable of deliberation, accountability, and restraint, guided by both conscience and the authority of command. The advent of autonomous systems fundamentally contests this premise rather than merely superficially. The trajectory IHL takes from this point forward does not depend on the regulatory timing. It is contingent on what kind of wars humanity chooses to

wage — and what kind of civilization it aspires to be, one that, although imperfectly, upholds the dignity of those who cannot fight.