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.