GC REAIM
GLOBAL COMMISSION ON RESPONSIBLE
AI IN THE MILITARY DOMAIN

Foreign, Commonwealth
& Development Office

GC REAIM Expert Policy Note Series

New and Emerging Terminologies in Ethical AI Principles: Exploring the International Law Implications in the Military Context

Keketso Kgomosotho

May 2025

POWERED BY
The Hague Centre
for Strategic Studies

GC REAIM Expert Policy Note Series

New and Emerging Terminologies in Ethical AI Principles: Exploring the International Law Implications in the Military Context

Authors: Keketso Kgomosotho

May 2025

Cover photo: Unsplash

The Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) is an initiative of the Government of the Netherlands that was launched during the 2023 REAIM Summit on Responsible Artificial Intelligence in the Military Domain in The Hague. Upon request of the Dutch Ministry of Foreign Affairs, the Hague Centre for Strategic Studies acts as the Secretariat of the Commission.

The GC REAIM Expert Policy Note Series was funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom. GC REAIM Experts maintained full discretion over the topics covered by the Policy Notes. The contents of the GC REAIM Expert Policy Note series do not represent the views of the Global Commission as a whole. The Policy Notes are intended to highlight key issues related to the governance of AI in the military domain and provide policy recommendations.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies
Email: info@hcss.nl
Website: www.hcss.nl

1. Introduction

The adoption of new terminologies in the military domain—whether in response to emerging technologies or evolving forms of violence—is not a novel phenomenon. For states and other duty-bearing actors, such linguistic shifts are neither incidental nor inconsequential, particularly when invoked to interpret or justify compliance with international legal obligations. For example, following the 9/11 attacks, states counterterrorism policies led to the introduction of terms such as "pre-emptive self-defence," "anticipatory self-defence," "elongated" or "expanded imminence," and the "unwilling or unable" doctrine, all aimed at reshaping international law on the use of force.[1] The invention of armed drones further accelerated this trend, giving rise to additional terminology like "signature strikes," "targeted killings," and the "global battlefield"—each attempting to rationalise new practices within the framework of existing international law.[2] Over two decades later, these terminologies continue to generate intense debate, with states deeply divided over their legitimacy, legal implications, and potential to erode foundational principles of international law.[3]

Equally, in the current attempts to establish comprehensive governance frameworks on AI – including in the military domain – there has been, yet again, the adoption and use of new terminologies that have serious implications for international law.[4] Such adoption is not random: First, this policy note cautions against uncritical adoption of emerging terms like "bias mitigation", "unintended engagements", "AI decision-making" which, in governance context, may be misaligned with established standards under international human rights law (IHRL) and international humanitarian law (IHL). Second, it warns stakeholders against the reinterpretation or misuse of legally defined terms such as

[1] Ashley S. Deeks, 'Consent to the Use of Force and International Law Supremacy', *Harvard International Law Journal* 54 (2013): 1–60; Monica Hakimi, 'Defensive Force against Non-State Actors: The State of Play', *International Law Studies* 91, no. 1 (15 January 2015), https://digital-commons.usnwc.edu/ils/vol91/iss1/1.

[2] Nils Melzer, *Targeted Killing in International Law* (OUP Oxford, 2008); Noam Lubell and Nathan Derejko, 'A Global Battlefield?: Drones and the Geographical Scope of Armed Conflict', *Journal of International Criminal Justice* 11, no. 1 (1 March 2013): 65–88, https://doi.org/10.1093/jicj/mqs096.

[3] Ruxandra Oana Vlad and John and Hardy, 'Signature Strikes and the Ethics of Targeted Killing', *International Journal of Intelligence and CounterIntelligence*, 2024, 1–29, https://doi.org/10.1080/08850607.2024.2382029; Michael Riepl, 'Can't Learn an Old Law New Tricks? Three Examples of How International Humanitarian Law Aged and Adapted', *Academy for European Human Rights Protection*, 30 January 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5117873; Rita Preto, 'A Never-Ending Tug-of-War: The Inherent Right of Self-Defense against Non-State Actors', *E-Publica* 11, no. 2 (30 July 2024): 32–58, https://doi.org/10.47345/v11n2art2; Jean Sikubwabo, 'A Critical Study of Legitimization of Preemptive Self-Defense as a Counter-Terrorism Measure Under International Law – WMO', 13 April 2020, https://worldmediation.org/a-critical-study-of-legitimization-of-preemptive-self-defense-as-a-counter-terrorism-measure-under-international-law/.

[4] Presentation of "Artificial Intelligence" as "Ai" with a lowercase "i" is a deliberate intellectual position to emphasise that these systems, while computationally sophisticated, do not demonstrate intelligence as traditionally understood or meaningfully theorised. The cognitive and intellectual capabilities attributed to these systems are fundamentally miscategorised as "intelligence," rather than simply exaggerated or misunderstood.

"commander" or "command responsibility" in ways that diverge from the authentic international criminal law (ICL). Just as states have critically examined the introduction of the term "meaningful human control" within the UN Group of Governmental Experts on Lethal Autonomous Weapon Systems (UN GGE on LAWS), they must apply the same level of scrutiny to other emerging terminologies that carry equally significant implications for existing international legal obligations. Finally, the policy note observes that while new terms such as "responsible AI" may be well-intended, they may be perceived by other actors as politically charged language and inadvertently undermine multilateral consensus.

1.1 Implications of New Terminologies for Procedural International Law

Because the introduction of new terminologies often reflects deliberate policy strategies by states, it is essential to critically assess their implications within the procedural framework of international law. International law's normative force derives fundamentally from State consent—a principle underlying both treaty formation and customary international law—establishing strict parameters for legal evolution as embodied in the VCLT[5] and reinforced through consistent State practice and *opinio juris*.[6] The *pacta sunt servanda* maxim[7] requires that modifications to international obligations occur through explicit State agreement or established customary law formation processes, a principle repeatedly affirmed by the ICJ in cases like *North Sea Continental Shelf*.[8] Treaty interpretation under VCLT Articles 31-32[5] provides limited scope for evolutionary interpretation, requiring terms be understood "in good faith in accordance with the ordinary meaning...in their context and in light of its object and purpose."[9] This interpretative framework is particularly stringent concerning jus cogens norms and erga omnes obligations,[10] where the ILC emphasizes that modifications require explicit State consent. Legitimate evolution of international legal obligations must: emerge from recognised sources enumerated in Article 38(1) of the ICJ Statute;[11] reflect clear State practice and *opinio juris* that is "sufficiently widespread, representative as well as consistent;"[12] and avoid undermining existing peremptory norms, such as the prohibition of indiscriminate attacks or the principle of non-discrimination.[13]

[5] *Vienna Convention on the Law of Treaties.*

[6] Nicaragua v. United States of America, Judgment, I.C.J. Reports 1986, 3, 98.

[7] I. I. Lukashuk, 'The Principle Pacta Sunt Servanda and the Nature of Obligation Under International Law', *American Journal of International Law* 83, no. 3 (July 1989): 513–18, https://doi.org/10.2307/2203309.

[8] See Article 26 of the Vienna Convention on the Law of Treaties; *North Sea Continental Shelf Cases*, I.C.J. Rep. 1969. See also *Nicaragua* ICJ Reps, 1986, p. 3 at 98, *Nuclear Weapons* and *Case of the SS Lotus* (1927).

[9] *Vienna Convention on the Law of Treaties, Article 31(1).*

[10] United Nations, *Report of the International Law Commission: Seventy-first Session (29 April–7 June and 8 July–9 August 2019)*, chap. 5, conclusion 23, *Peremptory Norms of General International Law (Jus Cogens)*, A/74/10, https://legal.un.org/ilc/reports/2019/english/a_74_10_advance.pdf.

[11] Statute of the International Court of Justice, Article 38.

[12] Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), Judgment, I.C.J. Reports 1986, 3, 98.

[13] Article 53, of the Vienna Convention on the Law of Treaties accordingly provides that a treaty will be void 'if, at the time of its conclusion, it conflicts with a peremptory norm of general international law. See also

Thus, introducing novel terminology in military AI ethics principles without satisfying these formal requirements risks creating parallel frameworks that potentially undermine the legitimacy and coherence of established legal standards,[14] which explains why similar terminological innovations in counterterrorism and drone warfare have been rejected by many states.

UNHRC Advisory Committee, 'A Global Call for Concrete Action for the Total Elimination of Racism, Racial Discrimination, Xenophobia and Related Intolerance and the Comprehensive Implementation of and Follow up to the Durban Declaration and Programme of Action' (23rd Session, 16 July 2019) A/HRC/AC/23/CRP.2 at 29.

[14] See historical examples in counterterrorism where terminological innovations have been scrutinized under international legal requirements.

2. "Meaningful Human Control" (MHC)

MHC has received extensive attention in multilateral discussions in the UN GGE on LAWS.[15] This policy note does not seek to rehash or provide an in-depth analysis of the concept itself, as that work is already the subject of considerable debate in UN *fora* and among states, and continues to evolve through diplomatic, academic, and technical discussions.[16] Rather, this section references MHC as a case study to illustrate how new terminologies introduced in the governance of emerging military technologies must be approached. The way MHC was initially introduced in relation to LAWS, and subsequently expanded to become a central concept in broader AI governance within the military domain, is a crucial point of reflection. The trajectory of this term—from a niche civil society conceptual tool to a central norm—demonstrates that the introduction of new language in international governance discourse must never be treated as a neutral or incidental act. New terminologies carry normative weight and interpretive consequences; their use can shape obligations, shift legal frameworks, and even redefine the standards by which state conduct is evaluated under international law. Most importantly, the level of scrutiny that MHC has received—with many states[17] and various

[15] United Nations, *Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems* (Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 25–29 March and 20–21 August 2019), CCW/GGE.1/2019/3, 17; International Committee of the Red Cross, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control* (Working paper submitted to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 20–21 August 2019), CCW/GGE.1/2019/WP.7; State of Palestine, *State of Palestine's Proposal for the Normative and Operational Framework on Autonomous Weapons Systems* (Working paper submitted to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 6–10 March and 15–19 May 2023), CCW/GGE.1/2023/WP.2/Rev.1; Thea Riebe, 'Meaningful Human Control of LAWS: The CCW-Debate and Its Implications for Value-Sensitive Design', in *Technology Assessment of Dual-Use ICTs: How to Assess Diffusion, Governance and Design*, ed. Thea Riebe (Wiesbaden: Springer Fachmedien, 2023), 111–32, https://doi.org/10.1007/978-3-658-41667-6_10.

[16] United Nations Institute for Disarmament Research, *Report of the Secretary-General: Lethal Autonomous Weapons Systems* (1 July 2024), UN Doc A/79/88; Paul Scharre, 'Autonomy, "Killer Robots," and Human Control in the Use of Force – Part II', *Just Security* (blog), 9 July 2014, https://www.justsecurity.org/12712/autonomy-killer-robots-human-control-force-part-ii/; United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*, UNIDIR Resources No. 2 (Geneva: UNIDIR, 2014), https://unidir.org/files/publication/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.

[17] See United Nations Institute for Disarmament Research, *Report of the Secretary-General: Lethal Autonomous Weapons Systems* (1 July 2024), UN Doc A/79/88, for an overview of states position on MHC in contexts of LAWS at pages 61–63, 94–97, 113–15; Austria, *Revised Working Paper* (Working paper submitted to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 6–10 March and 15–19 May 2023), CCW/GGE.1/2023/WP.1/Rev.1; Brazil, *Statement by Brazil, 78th UN General Assembly First Committee, 23 October 2023*, https://reachingcriticalwill.org/images/documents/Disarmament-fora/1com/1com23/statements/23Oct_Brazil.pdf; Türkiye, *Statement by Türkiye, Thematic Discussion on*

UN institutions[18] adopting it while other states sharply disagreeing[19]—should be regarded as an exemplary standard of multilateralism on such key language. This collective scrutiny affirms that all emerging terms in military AI governance with legal or normative implications must undergo a similarly rigorous process of critical assessment, legal evaluation, and state-led deliberation. Only through such processes can the integrity and coherence of international law be preserved in the face of rapid technological and linguistic evolution. The acceptability of new terminologies in international legal and governance frameworks cannot rest on the good intentions or benevolence of those introducing them; rather, it must be determined by the substantive implications such terms have for existing international legal obligations. The UN Human Rights Council, in its report on the human rights implications of AI in the military domain, explicitly cautioned against the uncritical adoption of new terminologies.[20]

"Conventional Weapons", First Committee, 77th Session of the United Nations General Assembly (21 October 2022).

[18] United Nations, *Lethal Autonomous Weapons Systems: Report of the Secretary-General*, A/79/88 (1 July 2024), https://digitallibrary.un.org/record/4059475?ln=en&v=pdf; United Nations Office for Disarmament Affairs, *Retaining Meaningful Human Control of Weapons Systems*, 16 October 2018, https://disarmament.unoda.org/update/retaining-meaningful-human-control-of-weapons-systems/.

[19] Sarah Knuckey, 'Governments Conclude First (Ever) Debate on Autonomous Weapons: What Happened and What's Next', *Just Security* (blog), 16 May 2014, https://www.justsecurity.org/10518/autonomous-weapons-intergovernmental-meeting/. For example, U.S. Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems*, January 25, 2023, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

[20] United Nations Human Rights Council Advisory Committee, *Report of the Human Rights Council Advisory Committee: Possible Impacts, Opportunities and Challenges of New and Emerging Digital Technologies with Regard to the Promotion and Protection of Human Rights*, UN Doc A/HRC/47/52 (19 May 2021), https://docs.un.org/en/A/HRC/47/52.

3. “Bias Mitigation” and “Minimise Unintended Bias”

The emergence of “bias mitigation” as a framework for addressing discriminatory outcomes in AI systems points to yet another problematic deviation from established international legal obligations, this time regarding non-discrimination. This is particularly concerning given non-discrimination's status as both *jus cogens* norm and *erga omnes* obligation.[21] As has been long established, non-discrimination constitutes a cornerstone of international law. Its *jus cogens* status reflects its fundamental importance to the international legal order. Under international treaty and customary law, prohibition of discrimination is absolute, admitting no derogation and imposing positive obligations on States to eliminate, not merely mitigate, discriminatory practices.[22] The emerging language of “mitigation of AI bias” in the AI governance discourse undermines the established legal framework of IHRL, which unequivocally demands the elimination or eradication of discrimination.

3.1 “Bias mitigation” is Inconsistent with IHRL

In their policy documents on AI in the military domain, various stakeholders are constantly using the term “mitigating bias” as an ethical principle that should be at the centre of AI governance.[23] Within the United Nations (“UN”) discussions on lethal autonomous weapon systems (“AWS”), reports have been submitted by States indicating a proposed policy to “*reduce unintended bias* in artificial intelligence capabilities relied upon in connection with the use of the weapon system.”[24] Among other things, they

[21] *RM & another v Attorney General* [2006] eKLR (Civil Case 1351 of 2002); (01 December 2006), page 25, available at http://kenyalaw.org/caselaw/cases/view/35204 wherein the High Court affirms that non-discrimination, discussed in the context of children, was “part of *jus cogen.”* The High Court relies here (at page 20) on the Human Rights Committee’s General Comment No. 18, which provides at para 1 that “non-discrimination constitutes a basic and general principle relating to the protection of human rights.” Inter-American Court of Human Rights, *"Mapiripán Massacre" v. Colombia*, Merits, Reparations and Costs; UN Human Rights Council, Report on Human rights implications of new and emerging technologies in the military domain (2024), para 19.

[22] Ibid. - United Nations, Report of the International Law Commission, Seventy-first session (29 April–7 June and 8 July–9 August 2019), chap. 5, conclusion 23, Peremptory norms of general international law (jus cogens), A/74/10 https://legal.un.org/ilc/reports/2019/english/a_74_10_advance.pdf

[23] See United Kingdom Ministry of Defence, *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence* (15 June 2022), p. 11, https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence; See United Nations, *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 (25 September 2019), annex IV, p. 13, https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

[24] Austria, Belgium, Canada, Costa Rica, Germany, Ireland, Luxembourg, Mexico, Panama, and Uruguay, *Addressing Bias in Autonomous Weapons*, Working paper submitted to the Group of Governmental Experts

recommend that States should take measures and safeguards aimed at mitigating risks such as "risk of unintended bias, such as on gender aspects and risk of unintended engagements."[25]

The introduction of "bias mitigation" terminology fundamentally alters this legal framework in several critical ways. First, it transforms an absolute prohibition into a matter of degree, suggesting that some level of discriminatory impact is acceptable if steps or efforts at mitigation are taken. This represents a fundamental departure from the absolute nature of non-discrimination obligations under international law - effectively weakening the normative force of non-discrimination requirements.

Second, it substitutes substantive obligations of results with procedural requirements (obligation of procedure or process). Where international law demands concrete outcomes—the elimination of discrimination—"bias mitigation" merely requires demonstrable steps or efforts at reduction.[26] This shift from outcome-based to process-based requirements fundamentally and qualitatively alter the nature of State and corporate obligations regarding discriminatory practices as articulated under international law.

Treaties such as the International Covenant on Civil and Political Rights ("ICCPR") and the International Convention on the Elimination of All Forms of Racial Discrimination ("ICERD") impose a binding obligation on States to prevent, prohibit, and eliminate all forms of discrimination, not merely to reduce its effects.[27] The provisions in these treaties "condemns" discrimination, "prohibit" discrimination, and demands "elimination" and "eradication" of discrimination in law and practice.[28]

Moreover, the shift from eliminating and eradicating discrimination to "mitigating AI bias" creates a lower threshold of accountability for States and private actors involved in AI development. Under IHRL, States are required to proactively dismantle systemic discrimination, ensure effective remedies for victims, and address the root causes of inequality. However, a governance approach centered on "bias mitigation" focuses primarily on symptom management rather than structural change, allowing

on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 4–8 March 2024, CCW/GGE.1/2024/WP.5, p. 6.

[25] As above, p.2.

[26] See for instance Article 10 of Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations.

[27] Article 2, Universal Declaration of Human Rights (UDHR) (1948); Articles 2(1) and 26, ICCPR (1966); Article 2(2), International Covenant on Economic, Social and Cultural Rights (ICESCR) (1966); Articles 2(1) and 5 of International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) (1965); Articles 2 and 5, Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) (1979); Article 2, Convention on the Rights of the Child (CRC) (1989); Article 5, Convention on the Rights of Persons with Disabilities (CRPD) (2006); Article 7, International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (ICMW) (1990); Article 2, Declaration on the Rights of Indigenous Peoples (UNDRIP) (2007).

[28] As above.

discriminatory AI systems to persist as long as they are perceived to be "less biased" than before, and as long as it can be demonstrated that steps have been taken to mitigate it. This is particularly concerning in the Global South, where AI-driven surveillance, predictive policing, and automated decision-making can cause disproportionate harm to populations which have historically experience discriminatory outcomes.[29] Instead of preventing harm at its root, the rhetoric of "bias mitigation" permits ongoing human rights violations under a veneer of progress, thereby undermining the non-derogable nature of the right to non-discrimination under international law.

Additionally, the language of "mitigating bias" equally dilutes the legal protections enshrined in regional human rights treaties such as the African Charter on Human and Peoples' Rights ("African Charter"), the European Convention on Human Rights ("ECHR"), and the American Convention on Human Rights ("ACHR"), all of which impose strict obligations on member States to eradicate discrimination in all forms.[30] Similarly, regional human rights courts and commissions such as the European Court of Human Rights ("ECHR") and the African Commission on Human and Peoples' Rights have developed jurisprudence emphasising that States must take positive measures to "eliminate discrimination", rather than merely reducing its impact.[31] If AI governance frameworks adopt a weaker standard of "bias mitigation," they risk undermining the legal force of international human rights treaties, allowing States and corporations to evade responsibility while continuing to deploy AI systems that perpetuate discriminatory and exclusionary outcomes

3.2 "Bias mitigation" is Inconsistent with IHL

The terminology of "mitigating AI bias" is further at odds with existing IHL, which establishes an absolute prohibition of discrimination in armed conflict, rather than a partial reduction of biased or discriminatory outcomes. The Geneva Conventions of 1949 and their Additional Protocols of 1977 enshrine the principle of non-discrimination as a fundamental component of the laws of war, requiring that all persons affected by armed conflict—whether civilians, prisoners of war, or wounded combatants—be treated without adverse distinction based on race, nationality, religion, or other protected

[29] Chinmayi Arun, 'AI and the Global South: Designing for Other Worlds', in *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale, and Sunit Das (Oxford University Press, 2020), https://doi.org/10.1093/oxfordhb/9780190067397.013.38.

[30] Articles 2, 3, and 18 (3), African Charter on Human and Peoples' Rights (ACHPR) (1981); Article 14, European Convention on Human Rights (ECHR) (1950); Articles 1(1) and 24, American Convention on Human Rights (ACHR) (1969).

[31] *Atala Riffo and Daughters v. Chile*, Judgment (Preliminary Objections, Merits, Reparations and Costs), Inter-Am. Ct. H.R. (ser. C) No. 239, February 24, 2012; *Centre for Minority Rights Development (Kenya) and Minority Rights Group International on behalf of Endorois Welfare Council v. Kenya*, Communication 276/03, African Commission on Human and Peoples' Rights (ACHPR), 2009; CJEU, C-33/89, Maria Kowalska v. Freie und Hansestadt Hamburg, 27 June 1990.

characteristics.[32] Equally, customary international humanitarian law prohibits discrimination on the grounds of race, gender, or any other prohibited ground.[33]

At the same time, stakeholders have noted that the application of AI technologies in decision-making – including AWS, intelligence surveillance, and targeting algorithms – raises profound concerns about the potential for discriminatory outcomes.[34] If these systems operate under a governance model that merely seeks to "mitigate bias" rather than eliminate discrimination, the risk of violating IHL norms becomes significantly heightened. In particular, the principle of distinction, a cornerstone of IHL, mandates that parties to a conflict must always distinguish between combatants and civilians, ensuring that civilians are never targeted.[35] AI systems used for military operations, if embedded with biased data or flawed algorithms, could wrongfully classify civilians as combatants, leading to unlawful targeting and violations of the prohibition on indiscriminate attacks. If AI governance only requires that such biases be mitigated rather than fully eliminated, there is no safeguard ensuring that lethal AWS or AI-assisted targeting systems comply with the strict non-discrimination requirements of IHL.

Similarly, the Geneva Conventions and customary IHL do not allow for partial compliance with non-discrimination rules; they impose a strict obligation on States and armed forces to ensure full adherence to the principle of equality in warfare. Furthermore, the Additional Protocols to the Geneva Conventions reinforce the absolute prohibition of discrimination by emphasising that all victims of war must receive equal protection and humane treatment, regardless of their status, nationality, or background. This extends to military detention, access to humanitarian aid, and the conduct of hostilities – all of which are increasingly subject to AI decision systems.[36] If AI governance frameworks normalises and legitimises the weaker standard of "bias mitigation," this could justify the

[32] Preamble, Articles 9(1), 10(1), and 75(1) of Additional Protocol I (1977) to the Geneva Conventions of 1949; Articles 2(1) and 4(1), Additional Protocol II (1977) to the Geneva Conventions of 1949; Common Article 3 to the Geneva Conventions of 1949; Articles 12 and 27, Geneva Convention I (For the Wounded and Sick in Armed Forces in the Field); Article 12, Geneva Convention II (For the Wounded, Sick, and Shipwrecked at Sea); Articles 13 and 16, Geneva Convention III (For Prisoners of War – POWs); Article 27, Geneva Convention IV (For the Protection of Civilians in Time of War).

[33] Rule 88, ICRC Study of Customary International Humanitarian Law.

[34] See Austria, Belgium, Canada, Costa Rica, Germany, Ireland, Luxembourg, Mexico, Panama, and Uruguay, *Addressing Bias in Autonomous Weapons*, Working paper submitted to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2024/WP.5 (8 March 2024), https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/CCW-GGE.1-2024-WP.5.pdf; United Nations Human Rights Council, *Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance: Artificial Intelligence*, UN Doc A/HRC/56/68 (2024), https://undocs.org/en/A/HRC/56/68, paras. 5–12 and 37–39.; United Nations, *Report of the Independent Expert on Human Rights and International Solidarity: Artificial Intelligence and International Solidarity – Towards Human-Centred Artificial Intelligence International Solidarity by Design*, UN Doc A/79/170 (2024), https://docs.un.org/en/A/79/170, paras. 5–18.

[35] International Committee of the Red Cross, "Principle of Distinction," *How Does Law Protect in War?*, https://casebook.icrc.org/law/principle-distinction.

[36] See Additional Protocol provisions above.

continued deployment of discriminatory AI-driven military systems that disproportionately impact certain populations—whether through predictive targeting, surveillance, or automated threat assessment. Such a shift would not only contradict existing treaty obligations under IHL but could also contribute to systematic violations of human rights in conflict zones, reinforcing global inequalities and allowing powerful States to deploy AI-driven warfare with reduced or by-passed accountability under international law.

Thus, the distinction between “eliminating discrimination” and “mitigating bias” is not mere semantics. It carries profound, qualitative legal and ethical implications for AI governance, particularly in the military domain. To uphold legal consistency and human rights protections, States and international institutions must insist on the language and legal standard of elimination rather than the performative language and standard of mitigation - ensuring that AI technologies are developed and deployed in full compliance with the *jus cogens* principles of equality, non-discrimination, and justice in international law.[37]

[37] T. Chengeta, ‘Autonomous Weapon Systems: Accountability Gap and Racial Oppression’, in *Reclaiming Human Rights in a Changing World Order*, ed. S. Christopher (London and Washington DC: Chatham House / Brookings Institution Press, 2022), 216–36, https://www.chathamhouse.org/sites/default/files/2022-10/2022-10-10-reclaiming-human-rights-changing-world-order.pdf; United Nations Human Rights Council Advisory Committee, *Human Rights Implications of New and Emerging Technologies in the Military Domain*, A/HRC/AC/33/CRP.1 (13 February 2025), https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/advisorycommittee/sessions/session33/neet-in-military-domain-a-hrc-ac-33-crp-1.docx, para. 19.

4. "Unintended engagements"

Similarly, in the discussions on targeting through AWS, a few States have introduced new terminologies such as “unintended engagements”, “unintended harm”, “unintended bias” and “minimisation of unintended engagements.[38] The term "unintended engagements" in military AI discourse appears designed to describe scenarios where AWS engage targets other than their intended objectives. However, a careful legal analysis reveals that such engagements would, in most if not all cases, constitute indiscriminate attacks which are already prohibited under international humanitarian law.

"Unintended engagements" typically encompass several categories of AWS behaviour. For instance, target misidentification, where an AI system incorrectly classifies a civilian object as a military objective,[39] or engagement spread where effects of an attack extend beyond the intended target, or system malfunction where technical failures lead to engagements outside predetermined parameters.

The United States DoD Directive defines “unintended engagements” as “the use of force against persons or objects that commanders or operators did not intend to be the targets of U.S. military operations, including unacceptable levels of collateral damage beyond those consistent with the law of war, ROE, and commander’s intent.”[40] The conduct described therein has in fact already received full treatment under established IHL, through the prohibition of indiscriminate attacks - another cornerstone of IHL, codified in Article 51(4)(a) of Additional Protocol I to the Geneva Conventions,[41] and also recognised as customary international law.[42] Similarly, this prohibition is absolute,

[38] See U.S. Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems*, January 25, 2023, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf; see also Austria, Belgium, Canada, Costa Rica, Germany, Ireland, Luxembourg, Mexico, Panama, and Uruguay, *Addressing Bias in Autonomous Weapons*, Working paper submitted to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 4–8 March 2024, CCW/GGE.1/2024/WP.5, 6; See Paul Scharre, ‘Autonomous Weapons and Operational Risk’ (Center for a New American Security, 2016), https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk Marta Bo, Laura Bruun, and Vincent Boulanin, ‘Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS’ (Stockholm International Peace Research Institute, 2022), p. 14-16, https://doi.org/10.55163/AHBC1664.

[39] For example, an autonomous system misidentifying a civilian vehicle as a military vehicle due to pattern recognition errors or an Ai systems confusing civilian gatherings with military formations due to similar heat signatures or movement patterns

[40] As above.

[41] According to Article 51(4)(a) of the 1977 Additional Protocol I, attacks “which are not directed at a specific military objective” and consequently “are of a nature to strike military objectives and civilians or civilian objects without distinction” are indiscriminate.

[42] International Committee of the Red Cross, “Indiscriminate Attacks,” *How Does Law Protect in War?*, https://casebook.icrc.org/a_to_z/glossary/indiscriminate-attacks.

admitting no exceptions or qualifications. In *Nuclear Weapons*, the ICJ has characterised it as one of the intransgressible principles of international customary law.[43]

4.1 "Unintended engagements" is Inconsistent with IHL

The Additional Protocol framework establishes clear criteria for what constitutes an indiscriminate attack. To that end, Article 51(4) defines indiscriminate attacks as those (a) which are not directed at a specific military objective; (b) which employ a method or means of combat which cannot be directed at a specific military objective; or (c) which employ a method or means of combat the effects of which cannot be limited as required by this Protocol.[44]

As such, each category of "unintended engagement" maps directly onto prohibited conduct under IHL, and a closer reading demonstrates that the emergent language of “unintended engagements” and the “minimisation of unintended engagements”[45] is fundamentally at odds, or inconsistent with the obligations set forth in IHL, particularly regarding the prohibition of indiscriminate attacks. IHL establishes a strict prohibition of indiscriminate attacks, which are not merely to be minimised but must be refrained from entirely.[46] The DoD's language, for instance, introduces a lower threshold of compliance by framing these engagements as unintended, which implicitly suggests that they are inevitable rather than unlawful acts that States must actively prevent. This divergence in terminology is legally significant, as it risks eroding the absolute nature of the IHL prohibition on indiscriminate attacks. The obligation under established IHL is not to minimise such attacks but to eliminate them entirely. The US approach, by merely seeking to reduce the probability of unintended engagements to “acceptable levels,” undermines the IHL requirement that indiscriminate attacks must never occur, creating room for and legitimising legally impermissible AI-driven military actions.

Furthermore, as indicated above, the DoD Directive’s defines “unintended engagements” to include attacks that cause incidental harm that is disproportionate to the military advantage gained. This too is inconsistent with the IHL principle of proportionality.[47] IHL explicitly prohibits any attack that is expected to cause excessive incidental loss of civilian life, injury to civilians, or damage to civilian objects relative to the anticipated military advantage.[48] In fact, under the IHL principle of proportionality, attacks that exceed this threshold are not merely unfortunate or unintended—they are considered to be

[43] ICJ, Nuclear Weapons Advisory Opinion (Para. 43). See also Israel, Operation Cast Lead (Part II, paras 120-126, 230-232, 365-392); Israel, The Targeted Killings Case (Paras 40-46); Israel, Human Rights Committee’s Report on Beit Hanoun (Para. 34, 38-42).

[44] Article 51(4), Additional Protocol I (1977) to the Geneva Conventions (1949).

[44] As above.

[45] U.S. Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems*, January 25, 2023, p. 23, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

[46] Article 51(4), Additional Protocol I (1977) to the Geneva Conventions (1949).

[47] U.S. Department of Defense, *DoD Directive 3000.09: Autonomy in Weapon Systems*, January 25, 2023, p. 23, https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf.

[48] Article 51(5)(b), Additional Protocol I (1977) to the Geneva Conventions (1949).

unlawful indiscriminate attacks.[49] Therefore, a policy that requires minimisation of such attacks directly contradicts IHL's categorical prohibition on disproportionate attacks, reinforcing the idea that compliance with IHL is not about reducing errors but about ensuring that certain forms of attacks never occur.

Equally concerning is the fact that the term "unintended engagements" is not found in any IHL treaty or customary IHL provisions. Instead, IHL uses legally established and precise terms such as "indiscriminate attacks," "excessive collateral damage," and "prohibited means and methods of warfare." This introduction of new, undefined terminology allows States to conveniently reinterpret established legal obligations in a way that dilutes their strength. When legally binding terms such as "prohibited" or "unlawful" are replaced with softer terms like "minimisation," the result is a gradual erosion of accountability. The continued use of non-IHL terminology in military AI governance will weaken international consensus on legal standards, making violations harder to define and enforce.

Once again, adherence to agreed-upon IHL terminology is not merely a matter of semantics; it is a critical mechanism for ensuring compliance and accountability in armed conflict. Terms such as "indiscriminate attacks" and "proportionality violations" carry clear legal implications and are backed by treaty provisions, judicial interpretations, and customary international law. Replacing these established terms with vague and malleable concepts like "unintended engagements" creates legal uncertainty and reduces the ability of victims to seek redress for unlawful harm caused by AI-driven military technologies. The international legal framework has been carefully developed to place absolute limits on conduct in warfare, and any deviation from agreed terminology risks diluting the protections provided to civilians and combatants alike.

As such, the DoD's – and other stakeholders' – approaches of "minimising unintended engagements" fail to align with IHL's clear and stringent requirements on the prohibition of indiscriminate attacks and the principle of proportionality. By substituting established legal prohibitions with language that implies mere reduction rather than elimination, stakeholders introduce a dangerous precedent that weakens IHL compliance in the context of military AI governance. States and international actors must resist such dilution and uphold the unequivocal IHL obligations that prohibit indiscriminate attacks, rather than simply seeking to mitigate their frequency or consequences.

4.2 "Unintended engagements" Versus the Concept of "mistake" under IHL

Moreover, a number of commentators, in defence of this emergent language, argue the new language of "unintended engagements" is not necessarily inconsistent with existing IHL because the concept of "mistake" is implicitly recognised, even if not explicitly

[49] As above.

defined, within IHL.[50] However, in the context of autonomous systems, it is crucial to distinguish between these terms. This paper must attend to this argument. As indicated above, the provided definition of "unintended engagements" refer to instances where force is used in a manner that results in harm beyond what was intended by commanders or operators, including collateral damage exceeding acceptable levels under the law of war. This concept is qualitatively different from a "mistake" as understood under IHL, particularly when considering mistake in the context of absolving criminal responsibility.

Under existing IHL, the assessment of a mistake is a qualitative, human-centric evaluation that examines factors such as reasonableness, adherence to precautionary measures, and the absence of bad faith (*malafides*).[51] A mistake that may absolve criminal liability must be one that a reasonable commander, acting in good faith and taking all feasible precautions, could have made under the circumstances. The standard is inherently tied to human attributes—judgment, situational awareness, moral agency, and the ability to reassess an evolving situation in real time.[52] These are qualities that Machine Learning (ML) systems lack, making any attempt to equate machine-driven errors with human mistakes legally and ethically flawed.

The notion of "unintended engagements" in the context of AWS thus introduces a mechanistic, probabilistic approach to the use of force, where errors are framed as a function of system limitations rather than violations of legal obligations. This is problematic under IHL because the law does not merely require minimising mistakes—it clearly prohibits indiscriminate attacks outright. Unlike human decision-makers, autonomous systems lack the ability to apply legal principles such as distinction and proportionality in the nuanced, context-sensitive manner required by IHL. A machine's failure to correctly identify a lawful target or reassess a situation mid-attack is not a legally recognisable "mistake" but rather an inherent limitation of delegating lethal decision-making to non-human entities.

The recharacterization of indiscriminate attacks as "unintended engagements" fundamentally alters the legal discourse through three critical dimensions. First, it inappropriately shifts focus from effects to intent, directly contradicting IHL's effects-based framework for evaluating the legality of attacks. Second, it transforms what are legally prohibited acts into technical incidents to be managed, effectively moving the discourse from legal prohibition to technical mitigation. Finally, as noted before, it

[50] Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict*, 3rd ed. (Cambridge: Cambridge University Press, 2016), para, 398 ("many things can go wrong in the execution of attacks, and, as a result, civilians are frequently harmed by accident."), https://doi.org/10.1017/CBO9781316389591.

[51] M. Schmitt and M. Schauss, 'Uncertainty in the Law of Targeting: Towards a Cognitive Framework', 2019, p.162, https://www.semanticscholar.org/paper/Uncertainty-in-the-law-of-targeting%3A-towards-a-Schmitt-Schauss/4e204cc07e394b66952780a08b6348600a962b38.

[52] As above, p.157.

suggests a relative standard based on technological capabilities, undermining the absolute nature of IHL prohibitions.

Ultimately, conflating human mistakes with machine limitations risks diluting the legal framework governing accountability in warfare. The assessment of a mistake under IHL hinges on human cognitive[53] and ethical faculties, and the introduction of AWS disrupts this foundation. Using the term "unintended engagements" to describe errors made by autonomous systems sidesteps the legal obligations of parties to an armed conflict, potentially eroding accountability under IHL. Rather than introducing vague new terminologies, it is critical to uphold existing IHL standards, which require that the use of force remains a human decision governed by legal and ethical principles—not a mathematical, empirical, statistical output of a ML algorithm.

[53] As above, p. 153, 157.

5. “AI decision-making”

5.1 “AI decision-making” under IHL

Another emerging term in the discourse on the use of force—adopted by stakeholders without a thorough examination of its implications for existing legal language and obligations under international law—is “AI decision-making.” Under IHL, decision-making on the use of force is not a singular event but a process that spans from the initiation of an attack against an adversary to its conclusion.[54] Under IHL, attacks are defined as “acts of violence against the adversary.”[55] This definition establishes that an attack encompasses the entire duration of violent actions taken against a legitimate target until such actions cease. The concept of an adversary under IHL refers to a party engaged in hostilities, whether in an international armed conflict (“IAC”), involving opposing State forces, or in a non-international armed conflict (“NIAC”), where the adversary may be a non-State armed group. The decision to use force against such an adversary involves these critical stages. First, the categorisation of a person or object as a lawful military target; next, the initiation of force against that designated adversary in accordance with other IHL rules such as proportionality; and finally, the cessation of force either after the neutralisation of the target or due to a change in the adversary’s legal status, requiring a reassessment of their legitimacy as a target.[56] Here, the fundamental question under IHL is whether non-human entities, such as ML-based autonomous systems, can be legally authorised to make such decisions? Applying treaty interpretation principles under international law, the answer is “no.”

The black-letter law of IHL, its historical development, and its core principles—including distinction, proportionality, and necessity—do not suggest that human decision-making during an attack can be preprogrammed or delegated to autonomous systems. The principles of IHL require continuous human value judgments and situational awareness in targeting decisions.[57] Distinction requires humans to determine whether a person or object is a lawful target; proportionality requires a human evaluation of collateral

[54] As above, p.149 (notes that targeting is a dynamic process characterised by situation-specific decision-making).

[55] Article 49(1) of Additional Protocol I (1977) to the Geneva Conventions (1949).

[56] Ingvild Bode, ‘The Problem of Algorithmic Bias and Military Applications of AI.’, *Humanitarian Law & Policy Blog* (blog), 14 March 2024, Ruben Stewart and Georgia Hinds, ‘Algorithms of War: The Use of Artificial Intelligence in Decision Making in Armed Conflict’, *Humanitarian Law & Policy Blog* (blog), 24 October 2023, https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/; Wen Zhou and Anna Rosalie Greipl, ‘AI in Military Decision-Making: Supporting Humans, Not Replacing Them’, *Humanitarian Law & Policy Blog* (blog), 29 August 2024, https://blogs.icrc.org/law-and-policy/2024/08/29/artificial-intelligence-in-military-decision-making-supporting-humans-not-replacing-them/.

[57] Thompson Chengeta, ‘Defining the Emerging Notion of “Meaningful Human Control” in Weapon Systems’, 2017, p. 871, https://www.semanticscholar.org/paper/DEFINING-THE-EMERGING-NOTION-OF-%E2%80%9C-MEANINGFUL-HUMAN-Chengeta/261db480a97725483c13bfd30836d9b6668a89e4.

damage in relation to military advantage; and necessity requires a judgment on whether force is justified under the circumstances. These are inherently human determinations that cannot be effectively pre-programmed or outsourced to statistical ML systems. Pre-programming or delegating such judgments to autonomous systems contradicts the purpose of IHL, which is to regulate human conduct in armed conflict by imposing moral, ethical, and legal constraints on human decision-makers.

A crucial aspect of human decision-making in targeting is the designation and continued designation of an individual or object as a lawful target throughout an attack. IHL does not permit an attack to proceed unchecked or indiscriminately. A target that was initially lawful may become unlawful due to a change in status or circumstances. For example, a combatant may become *hors de combat* (out of combat) by surrendering or being wounded, or a civilian object may lose its military significance. Such determinations require real-time human judgment, since ML-based autonomous system cannot assess changes in intent, context, legal status, or battlefield conditions with the nuanced reasoning required under IHL.[58] Further, if AI or other non-human systems were to make such determinations, they would lack the legal and ethical accountability necessary under IHL to ensure compliance with the law of armed conflict.

Emerging terminologies such as "AI decision-making" or claims that AI systems can "comply with IHL" risk fundamentally mischaracterising the legal framework governing targeting decisions. IHL is premised on the idea that obligations fall on human actors—States, commanders, and individual combatants—not ML algorithms. The legal responsibility for ensuring that force is used lawfully, proportionally, and discriminately rests with humans, not with autonomous systems. Introducing language that implies autonomous systems can "decide" or "comply with" IHL distorts legal obligations and creates a false narrative that non-human entities can bear responsibility under IHL. This is not merely a theoretical issue; it risks eroding legal accountability, as no algorithm can be held legally or morally responsible for war crimes. Therefore, adherence to established legal terminology is not just a matter of legal accuracy—it is essential to prevent the dilution of IHL and to maintain the fundamental principle that humans—not machines—must remain accountable for decisions to use force in armed conflict. The decision to use force is not a mechanistic process but a complex legal and moral determination requiring a reasoned application of IHL principles. It is not merely about selecting a target based on algorithmic parameters but about applying legal discretion, contextual interpretation, and accountability—elements that no ML system can perform effectively.

[58] As above, p.875M. Schmitt and M. Schauss, 'Uncertainty in the Law of Targeting: Towards a Cognitive Framework', 2019, p.152 (refers to multifaceted situational assessment when planning, approving or executing attacks), https://www.semanticscholar.org/paper/Uncertainty-in-the-law-of-targeting%3A-towards-a-Schmitt-Schauss/4e204cc07e394b66952780a08b6348600a962b38.

The authority to make decisions regarding the use of force is firmly vested in human agents—fighters and combatants—who bear responsibility for ensuring that attacks comply with IHL. The *lex lata* of IHL does not recognise non-human entities, such as autonomous systems, as lawful decision-makers in targeting. IHL explicitly assigns obligations such as verifying targets, taking precautions, and cancelling attacks if they become unlawful— to humans.[59] Only humans possess the legal capacity to make discretionary judgments, exercise moral reasoning, and be held accountable for violations of IHL. Granting AI systems the authority to "decide" to use force would contradict fundamental IHL principles by detaching targeting decisions from human responsibility and legal accountability.

It is thus a misnomer—even an inconsistency with IHL—to use terms such as "AI decision-making" or "autonomous weapon systems complying with IHL." Decision-making under IHL is not a purely empirical, computational process; it is a legal act that carries obligations and responsibilities that can, thus far, only be fulfilled by humans. Similarly, compliance with IHL is not merely about meeting algorithmic thresholds but about engaging in a process of legal reasoning, proportionality assessments, and ethical judgment, which an ML systems lack. The use of such misleading terminology risks diluting the legal framework of IHL by implying that machines can assume human obligations, when in fact, accountability and legal agency remain inseparably tied to human actors. The law of war is designed around the human exercise of discretion and responsibility—elements that AI, by its nature, cannot replicate.

5.2 "AI decision-making" under IHRL

Similarly, in the context of IHRL, the use of terms like "AI decision-making" and "AI systems complying with the law" introduces a conceptual and legal distortion. Under IHRL treaties, States are obligated to "respect, protect, fulfil, and promote human rights." These obligations are inherently State-driven and require human agents—governments, military officials, law enforcement, and policymakers—to ensure rights are upheld.[60] The existing legal framework does not recognise non-human entities as duty-bearers, for good reason. This means that compliance with human rights law cannot be assigned to AI systems or autonomous weapons.[61] If compliance with human rights obligations is framed in terms of AI decision-making, it shifts responsibility away from States and their human agents, weakening accountability mechanisms.

The fundamental premise of human rights law is that obligations are performed by human agents, not automated processes. The principles of due process, proportionality, and non-discrimination require reasoning, self-reflexivity, contextuality, interpretation,

[59] Article 57, Additional Protocol I (1977) to the Geneva Conventions (1949).

[60] Thompson Chengeta, 'Autonomous Weapon Systems and the Inadequacies of Existing Law: The Case for a New Treaty', *Journal of Law & Cyber Warfare* 8, no. 2 (2022): p. 111–124.

[61] As above.

and moral or value consideration—capacities that ML systems fundamentally lack, due to their exclusively quantitative, mathematical operational logic.[62] Moreover, IHRL establishes mechanisms for redress and accountability in cases of violations, requiring human actors to be held responsible for their actions. If an AI system makes a targeting decision that results in civilian harm or an unlawful use of force, it cannot be held accountable under human rights law in the same way a human commander or political authority can. The introduction of language that implies AI "compliance" with human rights obscures the necessity of human oversight and decision-making.

Beyond the legal misalignment, framing AI as a "decision-maker" in military and law enforcement contexts poses serious risks to human rights protections. If AI-driven systems are perceived as capable of making legally compliant decisions, there is a temptation by users to abdicate their responsibility to rigorously assess and review AI-based operations, leading to a dangerous erosion of oversight and accountability. This could result in arbitrary deprivations of life, algorithmic discrimination, and disproportionate uses of force, all contrary to IHRL's fundamental principles. The law is clear that only human actors are accountable for upholding and ensuring compliance with human rights—delegating such responsibilities to AI undermines the protective function of IHRL. For these reasons, States and international bodies must resist the adoption of misleading terminologies such as "AI decision-making" and "autonomous systems complying with the law." These phrases falsely suggest that legal obligations can be automated or mechanised, when in reality, human agency (and higher-order human cognitive capabilities) remains the cornerstone of both IHL and IHRL compliance. The language of international law must remain precise and human-centred, ensuring that legal and ethical responsibilities remain clearly attributed to States, military commanders, and decision-makers, rather than being diluted through technological abstraction.

[62] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep Learning', *Nature* 521, no. 7553 (May 2015): 436–44, https://doi.org/10.1038/nature14539.

6. Command Responsibility in AI Context

The final section examines why command responsibility, designed for human-to-human command relationships, fails to adequately address AWS deployment contexts, creating significant accountability gaps and potentially undermining the right to remedy. Specifically, this section examines how referring to individuals operating ML systems or tools as “commanders” is inconsistent with the provisions of IHL and international criminal law (“ICL”). In the currently evolving discourse on AI governance, the term “commander of AI systems” has been introduced, leading to distortions in the established meaning of "commander" under international law.

The doctrine of command responsibility, established through post-World War II jurisprudence and codified under IHL and ICL, rests on three essential elements. First, the existence of a superior-subordinate relationship with effective control; second, the superior's knowledge or constructive knowledge of subordinates' crimes; and finally failure to take necessary and reasonable measures to prevent or punish such conduct.[63] This framework presupposes specific characteristics of human command relationships. It presupposes a cognitive capacity for meaningful oversight, the ability to assess and influence subordinate behaviour, a shared understanding of legal and ethical obligations, and clear chains of command and control.

The doctrine of command responsibility is a sophisticated legal framework for attributing criminal responsibility to military commanders for the acts of their subordinates. The superior-subordinate relationship, in particular, requires demonstration of effective control—the material ability to prevent or punish criminal conduct. As such, we see that the term “commander” under IHL has a precise legal meaning, rooted in a human-to-human hierarchical relationship within military structures. IHL establishes the duty of commanders to prevent, suppress, and report breaches of IHL committed by persons under their command.[64] This duty presupposes a human superior-subordinate relationship in which the commander exercises direct and effective control over human forces. Autonomous systems, being non-human entities, do not possess agency, intent, or the capacity to be “commanded” in the way human subordinates are. Describing a human operator of a ML autonomous system as a “commander” distorts this well-established legal understanding and risks eroding the framework of accountability in military operations.

[63] *In re Yamashita*, 327 U.S. 1 (1946); International Committee of the Red Cross, “Command Responsibility,” *How Does Law Protect in War?*, https://casebook.icrc.org/a_to_z/glossary/command-responsibility;

[64] Article 87, Additional Protocol I (1977) to the Geneva Conventions (1949).

Furthermore, IHL explicitly requires commanders to ensure that subordinates are aware of their obligations under IHL.[65] This obligation presumes that subordinates are capable of comprehending (a semantic understanding), internalising, and executing legal and ethical directives—a capacity that ML systems categorically lack. Autonomous systems do not "understand" legal principles beyond statistical correlations; they operate based on pre-programmed parameters and statistical models (a mathematical formular, if you will). Thus, applying the term "commander" to a human interacting with a ML system fundamentally misrepresents the nature of command in IHL and risks weakening the enforcement of accountability mechanisms. Moreover, IHL places an obligation on commanders to take preventive and corrective actions when they become aware that subordinates may commit, or have committed, breaches of IHL.[66] This requirement presumes that subordinates operate with discretion and intent, which allows for their behaviour to be influenced or corrected by a commander. ML systems, however, do not, and cannot exercise independent judgment; their actions are determined by pre-programmed algorithms and machine learning processes.

6.1 Command Responsibility and ICL Concepts

Equally, under ICL, command responsibility applies strictly within a human-to-human relationship.[67] A "military commander or person effectively acting as a military commander" is criminally responsible for crimes committed by forces under their "effective command and control."[68] The core criterion for command responsibility is the ability of the commander to exercise control over subordinates, including preventing, repressing, and reporting crimes. ML-based AWS do not function as "forces" in the legal sense; they are tools that lack agency, legal personality, and the ability to form intent. Thus, using the term "commander" in relation to such systems is legally flawed and distorts established principles of criminal liability.[69]

Furthermore, ICL also requires that the commander "knew or, owing to the circumstances at the time, should have known" that subordinates were committing or about to commit crimes.[70] This presupposes that the commander is dealing with sentient individuals capable of making autonomous decisions, which is incompatible with the nature of AWS based on ML techniques. Such systems do not possess intent or moral culpability, meaning their actions cannot be equated with those of human subordinates.[71] Assigning the term "commander" to humans interacting with them

[65] As above.

[66] As above, Article 87(3).

[67] Article 28, Rome Statute of the International Criminal Court (ICC).

[68] As above.

[69] Thompson Chengeta, 'Accountability Gap, Autonomous Weapon Systems and Modes of Responsibility in International Law', *SSRN Electronic Journal*, 2015, https://doi.org/10.2139/ssrn.2755211.

[70] Article 28(1)(a), Rome Statute of the International Criminal Court (ICC).

[71] Thompson Chengeta, 'Accountability Gap, Autonomous Weapon Systems and Modes of Responsibility in International Law', *SSRN Electronic Journal*, 2015, https://doi.org/10.2139/ssrn.2755211.

creates a misleading narrative that complicates legal accountability and diminishes the effectiveness of ICL mechanisms.

Additionally, under ICL, command responsibility extends to situations where a superior has "effective authority and control" over subordinates.[72] This concept inherently depends on the ability of the commander to influence human actors through orders, training, and disciplinary measures. AWS, however, do not respond to disciplinary actions or commands in a legal sense. Rather than invoking command responsibility, the correct legal framework for assessing accountability in the deployment of AWS is individual criminal responsibility—whereby the human operator, programmer, or decision-maker may be held directly accountable for unlawful acts resulting from the use of the system. Retaining this distinction is critical to ensuring that legal responsibility remains human-centric and that machines are not erroneously treated as moral agents – a proposal that our current legal framework manifestly does not support.

[72] Article 28(2), Rome Statute of the International Criminal Court (ICC).

7. “Responsible AI”

While emerging terminologies such as “responsible AI” aim to foster good practices and do not inherently conflict with international law, they may nonetheless undermine established multilateral governance frameworks. Such language can appear politically charged, introducing unhelpful distinctions between ostensibly “responsible” and “irresponsible” actors, or between entities presumed to have good intentions and those assumed otherwise. The term “responsible AI” risks serving as a political façade, potentially facilitating ethics-washing and prioritising voluntary commitments over binding obligations under international law.

The term “responsible AI” emerged primarily from corporate and institutional narratives,[73] representing what might be characterised as “strategic regulatory pre-emption” rather than substantive governance. From a legal perspective, it raises fundamental definitional problems—lacking clear metrics or standards for what constitutes “responsible,” showing ambiguity about whether responsibility refers to development, deployment, or outcomes, and providing no clear mechanism for enforcement.[74] This terminological ambiguity serves as a strategic asset—its imprecision allows flexible interpretation while enabling actors to claim compliance without meeting specific legal standards, effectively shifting discourse from lex lata legal obligations toward ethical aspirations.[75] Unlike “responsible AI,” Article 36 of Additional Protocol I to the Geneva Conventions establishes clear, unambiguous legal obligations regarding weapons review,[76] requiring States to determine whether new weapons would be prohibited by international law based on specific, measurable criteria including compliance with explicit prohibitions under treaty law, adherence to customary principles, conformity with the Martens Clause,[77] and evaluation of indiscriminate

[73] See corporate initiatives from major technology companies facing scrutiny over AI deployments. See Thilo Hagendorff, ‘The Ethics of AI Ethics: An Evaluation of Guidelines’, *Minds and Machines* 30, no. 1 (1 March 2020): 99–120, https://doi.org/10.1007/s11023-020-09517-8.

[74] What is "responsible" in one context may not be in another, with no specificity regarding who determines what constitutes "responsible." See Anna Jobin, Marcello Ienca, and Effy Vayena, ‘The Global Landscape of AI Ethics Guidelines’, *Nature Machine Intelligence* 1, no. 9 (September 2019): p. 389 & 391 identifying over 80 Ai ethics documents with substantial divergence in their interpretation of principles, https://doi.org/10.1038/s42256-019-0088-2.

[75] This serves both corporate interests and State interests alike—where both seek fast adoption of AI technology, either because of profit imperatives or political (AI race) imperatives of military dominance. See for Elettra Bietti, ‘From Ethics Washing to Ethics Bashing: A View on Tech Ethics from within Moral Philosophy’, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20 (New York, NY, USA: Association for Computing Machinery, 2020), p. 210, https://doi.org/10.1145/3351095.3372860; Ben Wagner, ‘Ethics As An Escape From Regulation. From “Ethics-Washing” To Ethics-Shopping?’, in *Being Profiled: Cogitas Ergo Sum* (Amsterdam University Press, 2018), 84–89, https://www.degruyterbrill.com/document/doi/10.1515/9789048550180-016/html?lang=en.

[76] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 (Additional Protocol I), Article 36 provides that all new weapons must be capable of being used in compliance with lex lata IHL—treaty and customary.

[77] The Martens Clause appears in the preamble to the 1899 Hague Convention II and has been reaffirmed in subsequent IHL treaties, including Article 1(2) of Additional Protocol I; see also International Court of

effects under Article 51(4).[78] By substituting these precise legal requirements with ambiguous ethical aspirations, "responsible AI" transforms legally binding obligations into discretionary guidelines—undermining uniformity of legal obligations Article 36 was designed to ensure, lowering accountability thresholds in military operations, and complicating assessment of State compliance since it lacks specific benchmarks.[79] In practice, this rhetorical device risks legitimising practices that may be unlawful, for the sake of military dominance or profit maximisation.

Justice, Legality of the Threat or Use of Nuclear Weapons (Advisory Opinion) [1996] ICJ Rep 226, para 78. It provides that in cases not covered by specific international agreements, civilians and combatants remain under the protection of principles of international law derived from established custom, principles of humanity, and the dictates of public conscience.

[78] These requirements constitute clear legal obligations with standards for compliance, developed through widespread State practice and international jurisprudence. The ICRC's interpretative guidance emphasizes these reviews must be systematic, empirically based, and legally rigorous. See 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*, *International Review of the Red Cross* 88, no. 864 (December 2006): 931–956, https://international-review.icrc.org/sites/default/files/irrc_864_11.pdf.

[79] Where Article 36 requires specific legal assessments, "responsible AI" permits subjective interpretation of longstanding legal standards. See Kenneth Anderson, Daniel Reisner, and Matthew Waxman, 'Adapting the Law of Armed Conflict to Autonomous Weapon Systems', *Int'l L. Stud.* 90 (1 January 2014): 386.

8. Conclusions

This policy note has examined several emerging terminologies in the governance of AI technologies, particularly in the military domain. The uncritical transposition of AI-related terminologies from technical disciplines into international legal governance poses significant risks to the integrity of established international legal standards. While these terms are appropriate in technological contexts, where probabilistic and procedural approaches are acceptable, they are fundamentally incompatible with the substantive and absolute obligations of international law, particularly under international humanitarian law.

The key recommendation for States and stakeholders is to adhere to the agreed language enshrined in ratified treaties when discussing and formulating policies on AI in the military context. The introduction of new, ambiguous terminology complicates multilateral discussions, undermines legal clarity and certainty, and disrupts common understanding - ultimately hindering progress in AI governance.

The political reality is that these new terminologies in AI governance, such as “mitigating bias” and “unintended engagements,” are emerging from geopolitically powerful states that currently lead in the development of AI for military applications. These same States also dominate policy discussions and agenda-setting in the governance of AI in the military domain. As a result, these terms quickly gain traction, becoming the prevailing language in multilateral discussions without rigorous scrutiny of their implications for existing *lex lata*. Over time, they are presented as “agreed language,” despite the absence of broad, inclusive debate, or State consent as required by international law.

The history of international law and policymaking has long been marked by epistemic injustice, where language, terminology, and framing disproportionately reflect the interests of a select few powerful States rather than the global community.[80] To ensure a just and equitable approach to AI governance, stakeholders must resist the uncritical adoption of new terminologies that risk diluting or redefining established legal norms. Instead, they must insist on preserving language that reflects the binding obligations of international law, ensuring that policy developments serve all states rather than a privileged few. Moreover, this terminological shift cannot be separated from international competition for AI dominance, where states engaged in technological arms races have strategic interests in maintaining development flexibility while demonstrating only nominal compliance with international law, effectively facilitating continued AWS development despite potential conflicts with IHL obligations.

Moving forward, in my view, requires recognition that the challenges facing international legal frameworks extend beyond technical or doctrinal considerations, to encompass

[80] See Makau Mutua and Antony Anghie, ‘What Is TWAIL?’, *Proceedings of the Annual Meeting (American Society of International Law)* 94 (2000): 31–40.

broader political and economic dynamics. As such, the effective preservation of legal standards will require strengthened institutional mechanisms for evaluating new terminology, enhanced international cooperation to resist regulatory competition, development of economic incentives aligned with legal compliance; and the explicit rejection of technological determinism in legal evolution. The integrity of international law, and the protections it provides, must take precedence over both technological expedience and market imperatives.

About the Author

Mr. Keketso Kgomosotho

Keketso Kgomosotho is a technology attorney and research scholar whose work lies at the intersection of artificial intelligence, data, and law, with a focus on the fundamental tensions between machine learning decision systems and legal frameworks. Mr Kgomosotho is an Ars Iuris Doctoral Fellow at the University of Vienna, where his research examines the architectural capabilities and limitations of machine learning decision systems and their legal implications in regulated decision-making contexts. He also serves as a Research Fellow at the Institute for International and Comparative Law in Africa at the University of Pretoria.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies