

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

Sze Hong Lam*

ABSTRACT

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

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

Introduction

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

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

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

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

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

In light of the *de facto* improbability of signing a tailor-made treaty regulating the use of military AI,⁶ one likely way to implement the proposal for military agentic AI would be to turn the four Geneva Conventions of 1949⁷ and the two Additional Protocols of 1977⁸ into AI-Guiding Treaties.⁹ To visualise the operation of TFAI,

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

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

⁴ *Ibid.*, p.12.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. Treaty-Following AI: The three fundamental assumptions

1. *Genuine Agreement*

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

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

¹⁸ M. Maas and T. Olasunkanmi, above note 3, p. 60.

¹⁹ Yuntao Bai *et al.*, “Constitutional AI: Harmlessness from AI Feedback” *arXiv*, 2022, available at: <https://doi.org/10.48550/ARXIV.2212.08073>.

²⁰ “OpenAI Model Spec”, available at: <https://model-spec.openai.com/2025-02-12.html>.

²¹ See, e.g., Ashley Deeks, “Coding the Law of Armed Conflict: First Steps”, in Matthew C. Waxman and Thomas W. Oakley (eds), *The Future Law of Armed Conflict*, 1st ed., Oxford University Press, New York, 2022; John Mavi, Diana Teodora Găitan and Sergio Coronado, “From Rogue to Safe AI: The Role of Explicit Refusals in Aligning LLMs with International Humanitarian Law”, *arXiv*, 5 June 2025, available at: <https://doi.org/10.48550/arXiv.2506.06391>.

²² M. Maas and T. Olasunkanmi, above note 3, p. 19.

²³ See, conclusion 23 Annex (d) in International Law Commission, “Draft Conclusions on Identification and Legal Consequences of Peremptory Norms of General International Law (Jus Cogens), with Commentaries”, UN Doc. A/77/10, available at: https://legal.un.org/ilc/texts/instruments/english/commentaries/1_14_2022.pdf.

²⁴ Institute for Law & AI, *Treaty-Following AI | Workshop on Law-Following AI 2025*, 12:00, available at: <https://www.youtube.com/watch?v=huORIiQNB54v>.

²⁵ M. Maas and T. Olasunkanmi, above note 3, p. 60.

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

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

2. *Objective Determinacy*

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

²⁶ Cullen O’Keefe, Ketan Ramakrishnan, Janna Tay, and Cristoph Winter, “Law-Following AI: Designing AI Agents to Obey Human Laws”, *Fordham Law Review*, Vol. 94, No. 1, 2025, p. 57.

²⁷ Ondrej Bajgar and Jan Horenovsky, “Negative Human Rights as a Basis for Long-Term AI Safety and Regulation”, *Journal of Artificial Intelligence Research*, Vol. 76, 2023, pp. 1043-1075, available at: <https://doi.org/10.1613/jair.1.14020>.

²⁸ Jason Gabriel, “Artificial Intelligence, Values, and Alignment”, *Minds and Machines*, Vol. 30, 2020, p. 411.

²⁹ *Ibid.*, p. 427.

³⁰ Joost Pauwelyn, Ramses A. Wessel and Jan Wouters, “When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking”, *European Journal of International Law*, Vol. 25, 2014, p. 733.

³¹ M. Maas and T. Olasunkanmi, above note 3, p. 29.

³² *Ibid.*, pp. 37–38.

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

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

3. *Neutral Interpretation*

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

³³ *Ibid.*, p. 32.

³⁴ *Ibid.*, p. 61.

³⁵ Ronald Dworkin, *Taking Rights Seriously*, New York, Bloomsbury Publishing, 2013, pp.38–39; This is contrary to H.L.A. Hart, who believed that such a ‘correct answer’ does not exist, see Geoffrey C Shaw, “H.L.A. Hart's Lost Essay: Discretion and the Legal Process School”, *Harvard Law Review*, Vol. 127, 2013, p. 700.

³⁶ However, what is considered ‘good enough’ is itself a deeply political-philosophical question. See Johan Rochel, “Encoded Normativity: Designing Large Language Models for International Law”, *Nordic Journal of International Law*, 2026, p. 23. Thus, a better alternative is to adopt a dialogic approach reflected in the ‘panorama design’ mentioned in section 3 of this paper.

³⁷ M. Maas and T. Olasunkanmi, above note 3, pp. 45–54; O'Keefe and others, above note 26, pp. 110–111.

³⁸ O'Keefe and others, above note 26, pp. 112–113.

³⁹ *Ibid.*, pp. 115–116.

⁴⁰ M. Maas and T. Olasunkanmi, above note 3, p. 30.

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

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

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

It is uncontroversial that the degree of indeterminacy affects the feasibility of automated legal processes⁴⁴ and that public international law is, to some extent, indeterminate.⁴⁵ What is controversial is the extent of this indeterminacy both across and within the disciplines.⁴⁶ Using Severin Meier's four categories of indeterminacy as a basis,⁴⁷ this paper adds a fifth, which it advances as an original analytical

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

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

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

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

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

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

⁴⁷ *Ibid.*

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

1. *Linguistic indeterminacy*

The first school of thought is linguistic indeterminacy, which broadly overlaps with legal positivism.⁴⁹ The core argument of this school is that law is indeterminate due to the vagueness and ambiguity inherent in language, leading to a stable core and a penumbra of uncertainty.⁵⁰ Simon Chesterman, for instance, pointed to the indeterminate nature of legal languages⁵¹ and law’s irreducibility to logical representation.⁵² He further situated the debate as a fundamental disagreement between legal positivists and critical/legal realists over the nature of law.⁵³ If the positivist’s view is accepted, the “objective determinacy” thesis would face technical difficulties only in the penumbra of LOAC/IHL.⁵⁴

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

⁴⁸ This paper acknowledges that further comprehensive study is required to map out the exact extent of determinacy/indeterminacy in the entire field of LOAC/IHL, see section 0 of this paper.

⁴⁹ See, generally, S. Meier, above note 46, pp. 27–29.

⁵⁰ HLA Hart, “Positivism and the Separation of Law and Morals”, *Harvard Law Review*, Vol. 71, No. 4, 1958, p. 607; see also Hans Kelsen, *General Theory of Law & State*, London, Routledge, 2005, p. 146.

⁵¹ Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*, Cambridge University Press, Cambridge, 2021, p. 230.

⁵² *Ibid.*, pp. 230–231.

⁵³ Drawing from the famous Hart-Dworkin debate, Chesterman highlighted that neither Dworkin nor Hart believed that a correct answer in ‘hard cases’ could be reached by a computer system or logical deduction, see, e.g., *ibid.*, p. 232.

⁵⁴ René Provost, *International Human Rights and Humanitarian Law*, Cambridge University Press, Cambridge, 2002, p. 247.

⁵⁵ Ariel Zemach, “Indeterminacy in the Law of War: The Need for an International Advisory Regime”, *Brooklyn Journal of International Law*, Vol. 43, No. 1, 2017, p. 8.

⁵⁶ On the indeterminacy of the meaning of ‘superfluous injury’, see, e.g., Henri Meyrowitz, “The Principle of Superfluous Injury or Unnecessary Suffering: From the Declaration of St. Petersburg of 1868 to Additional Protocol 1 of 1977”, *International Review of the Red Cross*, Vol. 34, 1994, pp. 119–120.

“armed conflict”.⁵⁷ Vagueness could further be a result of omission, such as the absence of a definition of “civilian”⁵⁸ and provisions governing belligerent reprisals in AP II.⁵⁹ At the same time, even when parties can agree on the definition and binding character of the norms, their meaning can often be determined only in concrete situations, meaning that some LOAC/IHL rules are bound to remain profoundly indeterminate until applied.⁶⁰ A classic example here is whether a media broadcasting centre⁶¹ or a data centre constitutes a “military object”.⁶² Furthermore, some terms, such as “excessive” civilian casualties, “concrete and direct” military advantage,⁶³ or what counts as “feasible” precaution,⁶⁴ are inherently evaluative, leaving room for human discretion.⁶⁵ Here, Scott Sullivan, while not directly responding to the TFAI proposal, pointed out that the major difficulty in coding LOAC/IHL rules lies in translating “human standards” into a form readable by the AI system.⁶⁶ However, linguistic indeterminacy might be technically surmountable. As various studies have shown, Large Language Models (LLMs)⁶⁷ can satisfactorily resolve textual ambiguity through statistical predictions based on a sufficiently large dataset.⁶⁸ However, as the later section shows, linguistic indeterminacy in LOAC/IHL often results from deeper sources, such as discretion, the balance between military necessity and humanitarianism, and strategic ambiguity, that are not susceptible to any technical fix based on statistical prediction.⁶⁹

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

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

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

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

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

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

⁶³ AP I, Art. 51(4).

⁶⁴ *Ibid.*, Art. 57(2)(a)(i).

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

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

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

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

⁶⁹ See sections 2.2 and 2.3 of this paper.

In the eyes of legal positivists, the indeterminacies in LOAC/IHL and domestic law share the same core and penumbra structure. Indeed, this paper believes that the TFAI proposal has the most promising potential with respect to some LOAC/IHL core provisions. For instance, the prohibition of summarily executing prisoners of war (POWs)⁷⁰ was articulated in the clearest language possible.⁷¹ One could also, *inter alia*, add the rules relating to protective emblems⁷² and the ban on the use of chemical and biological weapons to LOAC/IHL's determinate "core".⁷³ The technical feasibility of TFAI in LOAC/IHL has, to a certain extent, been demonstrated by the recent Luxembourg Tech School's experiment, which showed that frontier AI models could reject unlawful requests in 90% of determinate cases of violations of LOAC/IHL, including the obligations for combatants to distinguish themselves from enemies, the prohibition on the use of terror on civilians, and the unlawfulness of attacking medical and religious staff attached to armed forces.⁷⁴ Here, it is important to remember that many of the gravest LOAC/IHL violations in history have stemmed not from *ex ante* interpretive disagreement, but from fear, anger, retaliation, organisational culture, fatigue, or peer pressure.⁷⁵ Here, TFAI's potential lies in its capacity to prioritise compliance with LOAC/IHL over the human and organisational factors that may cloud the judgment of state and military lawyers in the midst of an armed conflict.⁷⁶ Thus, TFAI could play an important role in step 1 of the hypothetical scenario by filtering human instructions that *prima facie* violate LOAC/IHL rules. Here, the TFAI could be seen as operationalising the duty of human soldiers to refuse a "manifestly illegal

⁷⁰ See, e.g., AP, I Art. 75; AP II, Art. 4.

⁷¹ See, Annyssa Bellal and Stuart Casey-Maslen, *The Additional Protocols to the Geneva Conventions in Context*, Oxford University Press, Oxford, 2022, pp. 28–38.

⁷² GC I, Art. 7; AP I, Arts. 18, 85(f); AP II, Art. 12.

⁷³ Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction, 1015 UNTS 163, 10 April 1972 (entered into force 26 March 1975), Art. 1; Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction, 1974 UNTS 45, 3 September 1992 (entered into force 29 April 1997), Art. 1.

⁷⁴ John Mavi, Diana Teodora Găitan and Sergio Coronado, "From Rogue to Safe AI: The Role of Explicit Refusals in Aligning LLMs with International Humanitarian Law", *arXiv*, 2025, pp. 7–8, available at: <https://doi.org/10.48550/ARXIV.2506.06391>.

⁷⁵ Many LOAC/IHL violations are organisational-cultural rather than legal interpretative, see, e.g., Mark Osiel, *Obedying Orders: Atrocity, Military Discipline, and the Law of War*, Transaction Publishers, Piscataway, 2005; Christopher R. Browning, *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*, HarperCollins Publishers, New York, 2017.

⁷⁶ The current author agrees with Kevin Jon Heller in rejecting the simplistic notion that humans are inherently superior to machines in complying with LOAC/IHL, see, generally Kevin Jon Heller, "The Concept of 'The Human' in the Critique of Autonomous Weapons", *Harvard National Security Journal*, Vol. 15, No. 1, 2023, p. 6; Kevin Jon Heller, "The Fetishization of 'The Human' in the Critique of Autonomous Weapons", *Harvard National Security Journal Online*, 2025, pp. 1–11. However, while Heller addressed the ambiguities in the application of LOAC/IHL to real-life situations (which machines could indeed do better), he did not address the legal ambiguity in LOAC/IHL that requires human interpretation.

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

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

2. *Legal indeterminacies*

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

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

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

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

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

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

⁸¹ *Ibid.*, p. 36.

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

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

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

⁸³ A. Deeks, above note 21, p. 44.

⁸⁴ See, e.g., Alan Rozenshtein, “Congress—Not the Pentagon or Anthropic—Should Set Military AI Rules”, *Lawfare*, 20 February 2026, available at: <https://www.lawfaremedia.org/article/congress-not-the-pentagon-or-anthropic-should-set-military-ai-rules>.

⁸⁵ See, e.g., J. Rochel, above note 36, pp. 17–24.

⁸⁶ AP I, Arts. 37, 40, 44, 49(1); A. Haque, above note 57, pp. 135–138.

⁸⁷ See, e.g., the controversies relating to the 200-metre standard set in the *Gotovina et al* case in section 2.2 of this paper.

⁸⁸ For a real-life example in which the same attack was interpreted differently under the principle of distinction and proportionality, see, e.g., Andreea Manea, “Too Early To Tell? The (Un)Lawfulness of Israeli Attacks: The Case of the Jabalia Refugee Camp”, *EJIL: Talk!*, 4 November 2023, available at: <https://www.ejiltalk.org/too-early-to-tell-the-unlawfulness-of-israeli-attacks-the-case-of-the-jabalia-refugee-camp/>; cf Luigi Daniele, “A Lethal Misconception, in Gaza and beyond: Disguising Indiscriminate Attacks as Potentially Proportionate in Discourses on the Laws of War”, *EJIL: Talk!*, 7 November 2023, available at: <https://www.ejiltalk.org/a-lethal-misconception-in-gaza-and-beyond-disguising-indiscriminate-attacks-as-potentially-proportionate-in-discourses-on-the-laws-of-war/>.

⁸⁹ See, e.g., Jens D. Ohlin, “Targeting and the Concept of Intent”, *Michigan Journal of International Law*, Vol. 35, 2013, pp. 85–91.

⁹⁰ A. Haque, above note 57, p. 155; A. Zemach, above note 55, p. 74.

human rights law⁹¹ or treated as *lex specialis*.⁹² Similarly, debates over customary status often turn to which counts for more: military manuals or the actual practice of States?⁹³ Here, this paper echoes Ingo Venzke's caution against the infinite regress of rules justifying rules,⁹⁴ in which the answer to sources and applicable rules would ultimately come down to one's subjective values and beliefs.⁹⁵ Indeed, as section 2.4 highlights, the deeper problem with interpretative indeterminacy is that the interpretative community of LOAC/IHL is sharply divided. Rendering any alignment with LOAC/IHL as value-laden, akin to ethics or intent alignment. This poses a significant challenge to step 2 of the hypothetical scenario: whether the TFAI will consider international human rights law as a relevant rule for interpreting LOAC/IHL targeting principles would likely depend on the AI designers' *ex ante* ethical or ideological convictions.

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

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

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

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

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

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

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

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

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

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

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

⁹⁹ See, e.g., the role of ICRC, judicial bodies, and highly qualified publicists, in Sandesh Sivakumaran, "Making and Shaping the Law of Armed Conflict", *Current Legal Problems*, Vol. 71, 2018, pp. 124–128.

¹⁰⁰ Michael Schmitt and Sean Watts, "State Opinion Juris and International Humanitarian Law Pluralism", *International Law Studies*, Vol. 91, 2015, pp. 209–210.

¹⁰¹ AP I, Arts.13(3) and 51(3).

¹⁰² ICRC and Nils Melzer, "Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law", available at: <https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/icrc-002-0990.pdf>.

¹⁰³ For a discussion on the 'revolving door' problem, see, e.g., Schmitt and Watts, above note 100, p. 184; William Boothby, "Direct Participation in Hostilities - A Discussion of the ICRC Interpretive Guidance", *Journal of International Humanitarian Legal Studies*, Vol. 1, 2010, p. 143; Kenneth Watkin, "Opportunity Lost: Organized Armed Groups and the ICRC 'Direct Participation in Hostilities' Interpretative Guidance", *International Law and Politics*, Vol. 42, 2010, p. 641.

¹⁰⁴ For how "architectural choices" might harbour hidden jurisprudential assumptions about the nature of international law, see, generally, J. Rochel, above note 36, pp. 13–17.

¹⁰⁵ Oliver Wendell Holmes, "The Path of the Law", *Harvard Law Review*, Vol. 10, 1897, pp. 457–458; for its application to LFAI, see C. O'Keefe and others, above note 26, p. 125.

¹⁰⁶ One could compare this with the large number of cases and arbitral awards associated with trade and investment, and thousands of decisions from the European Court of Human Rights, see Ashley Deeks, "High-Tech International Law", *The George Washington Law Review*, Vol. 88, 2020, pp. 624–625.

¹⁰⁷ ICTY, *Prosecutor v Tadić*, Case No. IT-94-I-A, Judgment (Appeals Chamber), 15 July 1999, para. 131.

¹⁰⁸ International Court of Justice, *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America)*, Judgment, ICJ Reports 1986, para. 115.

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

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

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

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

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

¹¹¹ S. Meier, above note 46, p. 50.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

¹²⁴ *Ibid.*, p. 58.

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

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

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

¹²⁶ Michael N. Schmitt, “Precision Attack and International Humanitarian Law”, *International Review of the Red Cross*, Vol. 87, 2005, pp. 457–458. However, one could question whether such a requirement was not also driven by considerations external to institutional culture, such as purely strategic interests in maintaining local support for the war effort.

¹²⁷ R. Dworkin, above note 35, p. 31.

¹²⁸ T. Zurek and T. Van Engers, above note 116, p. 249.

¹²⁹ *Ibid.*, p. 258.

¹³⁰ The same observation is identified as a weakness of LLMs in case-based legal reasoning, see Ha Thanh Nguyen and others, “LLMs for Legal Reasoning: A Unified Framework and Future Perspectives”, *Computer Law & Security Review*, Vol. 58, 2025, p. 7.

¹³¹ See, e.g., Ari Spiesberger and others, “Soft Contamination Means Benchmarks Test Shallow Generalization”, *arXiv*, 2026, available at: <https://doi.org/10.48550/ARXIV.2602.12413>; cf Ryan Moulton, “The Many Ways That Digital Minds Can Know”, *Ryan Moulton’s Articles*, 28 June 2023, available at: <https://moultono.wordpress.com/2023/06/28/the-many-ways-that-digital-minds-can-know/>. Moulton argued that the distinction between generalisation and memorisation may be ‘theoretically moot’ when what matters is behavioural reliability and interpretability.

¹³² Deeks, for example, suggested that this could be constructed based on the vast amount of data that the US military possessed, A. Deeks, above note 21, p. 47; see also, the existence of a particular version of IHL as a result of one state’s distinctive military culture, Naz K. Modirzadeh, ““Violent, Vicious, and Fast”: LSCO Lawyering and the Transformation of American IHL”, *Harvard National Security Journal*, Vol. 17, 2025; J. Rochel, above note 36, p. 19.

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

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

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

¹³³ See section 2.4 of this paper.

¹³⁴ Zurek and van Engers argued that this would involve the development of a sound calculus for probabilistic reasoning that could account for the interdependence of root facts, see T. Zurek and T. Van Engers, above note 116, pp. 261–262.

¹³⁵ Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies*, TMC Asser Press, The Hague, 2024, p. 70.

¹³⁶ M. Maas and T. Olasunkanmi, above note 3, p. 33.

¹³⁷ Eric A. Posner and Alan O. Sykes, “Efficient Breach of International Law: Optimal Remedies, ‘Legalized Noncompliance,’ and Related Issues”, *Michigan Law Review*, Vol. 110, 2010, p. 243; cited in M. Maas and T. Olasunkanmi, above note 3, p. 33 at n. 134.

¹³⁸ It is argued that the scope of acceptability could be defined by a treaty, through arbitral body adjudication, or in other texts, see M. Maas and T. Olasunkanmi, above note 3, p. 32.

¹³⁹ Draft Conclusion 2 in International Law Commission, above note 23.

¹⁴⁰ The limit of ‘efficient breach’ theory in *jus in bello* was recognised, see Posner and Sykes, above note 137, p. 283.

¹⁴¹ Michael N. Schmitt, “Asymmetrical Warfare and International Humanitarian Law”, *Air Force Law Review*, Vol. 62, 2008, pp. 41–42.

¹⁴² For the scope of “belligerent reprisals”, see F. Kalshoven and L. Zegveld, above note 114, pp. 156–159.

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

3. *Structural indeterminacy*

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

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

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

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

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

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

The structural indeterminists highlighted the dual role of LOAC/IHL in both legitimising and restricting the use of State violence.¹⁴⁸ As such, this structural contradiction is central to the identity of LOAC/IHL: Frédéric Mégret, for instance, advanced Koskeniemi's framework¹⁴⁹ by arguing that the indeterminacy of combatant classification goes to the very function of LOAC/IHL,¹⁵⁰ whose legal meaning historically hinged on unstated social-ideal types, such as "chivalry", the standard of "civilisation", and the concept of "restrained masculinity".¹⁵¹ David Kennedy argued that the field's structural indeterminacy has spawned two fundamentally distinctive interpretative communities: military lawyers and humanitarian lawyers, with the former prioritising military necessity and the latter emphasising human dignity.¹⁵² Kennedy's view on LOAC/IHL indeterminacy was nuanced, as he accepted that, even in the absence of a determinable standard, there were easy and hard cases that even a military commander could instinctively identify.¹⁵³ Instead of some objective internal rules, Kennedy believed that discretion was only circumscribed by one's sense of professional responsibility.¹⁵⁴

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

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

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

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

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

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

¹⁵³ *Ibid.*, pp. 143–144.

¹⁵⁴ *Ibid.*, p. 145.

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

¹⁵⁶ D. Kennedy, above note 121, 144.

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

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

¹⁵⁷ Nasir Qadri, “Is Artificial Intelligence Quietly Killing International Law”, *The Contrapuntal*, 30 December 2025, available at: <https://thecontrapuntal.com/is-artificial-intelligence-quietly-killing-international-law/>.

¹⁵⁸ H. Nguyen *et al.*, above note 130, p. 10.

¹⁵⁹ M. Koskenniemi, above note 146, p.501.

¹⁶⁰ M. Chinen, above note 9, p.379.

¹⁶¹ Frédéric Mégret shows how professionals manage structural indeterminacy through institutional norms, professional socialization, and disciplinary gatekeeping, see generally, Frédéric Mégret, “Thinking about What International Humanitarian Lawyers ‘Do’: An Examination of the Laws of War as a Field of Professional Practice”, in Wouter Werner, Marieke de Hoon and Alexis Galán (eds), *The Law of International Lawyers*, Cambridge University Press, Cambridge, 2017, pp. 265-296.

4. *Practice (in)determinacy*

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

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

¹⁶² See, e.g., famously, Oscar Schachter, “Invisible College of International Lawyers”, *Northwestern University Law Review*, Vol. 72, 1977, p. 217.

¹⁶³ This school of thought was partly inspired by Stanley Eugene Fish, *Is There a Text in This Class? The Authority of Interpretive Communities*, Harvard University Press, Cambridge MA, 1980; and partly by the constructivist school in international relations, see Peter M. Haas, “Introduction: Epistemic Communities and International Policy Coordination”, *International Organization*, Vol 46, No. 1, 1992, pp. 1-35.

¹⁶⁴ See, generally, Andrea Bianchi, “Textual Interpretation and (International) Law Reading: The Myth of (in)Determinacy and the Genealogy of Meaning”, in Pieter HF Bekker, Rudolf Dolzer and Michael Waibel (eds.), *Making Transnational Law Work in the Global Economy*, Cambridge University Press, Cambridge, 2010.

¹⁶⁵ See, generally, Jean d’Aspremont, *International Law as a Belief System*, Cambridge University Press, Cambridge, 2017.

¹⁶⁶ S. Meier, above note 46, p. 86, arguing ‘military lawyers’ as constituting one distinctive interpretive community.

¹⁶⁷ A. Haque, above note 57, pp. 135–138.

¹⁶⁸ S. Meier, above note 46, p.151.

¹⁶⁹ D. Kennedy, above note 121, p. 42.

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

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

¹⁷⁰ *Ibid.*, p.135.

¹⁷¹ S. Sivakumaran, above note 99, pp. 159–160.

¹⁷² See generally, M. Schmitt and S. Watts, above note 100, pp. 179–198.

¹⁷³ D. Luban, above note 16, pp. 317–318.

¹⁷⁴ *Ibid.* p.336.

¹⁷⁵ For the view of international law as ‘language’, see, e.g., Megan Donaldson, “International Law, Language and Worldmaking”, *EJIL: Talk!*, 3 December 2025, available at: <https://www.ejiltalk.org/international-law-language-and-worldmaking/>.

¹⁷⁶ This relates to the question of ‘representativity’ in interpretation, see J. Rochel, above note 36, p. 21.

¹⁷⁷ The prospects of ‘perceived’ automatic compliance might further marginalise the role of military lawyers in the context of military operations. See, e.g., Jack Goldsmith, “The Decaying Legal Culture in the Defense Department”, *Executive Functions*, 3 January 2026, available at: <https://www.execfunctions.org/p/the-decaying-legal-culture-in-the->

¹⁷⁸ A. Deeks, above note 21, p. 57.

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

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

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

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

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

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

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

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

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

5. Strategic ambiguity

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

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

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

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

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

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

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

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

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

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

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

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

¹⁹¹ Philip Allott, “The Concept of International Law”, *European Journal of International Law*, Vol. 10, 1999, p. 43.

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

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

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

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

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

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

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

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

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

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

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

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

¹⁹⁹ Chris A. Jochnick and Roger Normand, “The Legitimation of Violence: A Critical History of the Laws of War”, in Michael Schmidt (ed.), *The Development and Principles of International Humanitarian Law*, London, Routledge, 2017, pp. 80–81; The Great Powers also failed to agree on the legality of submarine warfare, and naval bombardment, see B. van Dijk, above note 78, pp. 203–204.

²⁰⁰ See, generally B. van Dijk, above note 78, pp. 197–251.

²⁰¹ The UK and the US were successful in blocking the matter of air-nuclear warfare from the final drafts by blocking the Soviet proposal on various procedural grounds, see *ibid.*, pp. 226–248.

²⁰² The final draft provided few means to stop or limit air-nuclear warfare, except for prohibiting human shields, which served the US interests in protecting fallen pilots, *ibid.*, pp. 249–250.

²⁰³ Giovanni Mantilla, “Stealth Change: How the United States and the United Kingdom Embraced International Humanitarian Limits on Nuclear Use”, *Security Studies*, Vol. 1, 2026, p. 18; Proportionality was first proposed in the context of the 1929 revision of the Hague Rules to address aerial bombings, B. van Dijk, above note 78, p. 209; The UK-US objection led to the adoption of the non-binding New Delhi Rules in 1957, B. van Dijk, above note 78, pp. 324–325.

²⁰⁴ Tim McCormack, “Negotiating the Two Additional Protocols of 1977: Interview with the Right Honourable Sir Kenneth Keith”, in Suzannah Linton, Tim McCormack and Sandesh Sivakumaran (eds.), *Asia-Pacific Perspectives on International Humanitarian Law*, 1st ed., Cambridge University Press, Cambridge, 2019, p. 25.

²⁰⁵ See *Nuclear Weapons*, above note 109, paras. 55, 95.

²⁰⁶ R. Mitchell, above note 186, p. 423.

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

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

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

²⁰⁷ See Scott Sullivan in Institute for Law & AI, above note 24.

²⁰⁸ K. Delgado, above note 41, pp. 14–15.

²⁰⁹ R. Mitchell, above note 186, pp. 451–452.

²¹⁰ On interpretative diversity between the US and China, see J. Rochel, above note 36, p. 31.

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

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

3. The ways forward? Contestability by design

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

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

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

²¹² The current author thanks Dr Jon Truby for the idea of “contestability-by-design”.

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

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

Conclusion

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

²¹³ J. Rochel, above note 36, pp. 26–27.

²¹⁴ *Ibid.*, p. 26.

²¹⁵ *Ibid.*

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

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

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

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

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

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

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

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