Beyond a Piecemeal Approach: Prospects for a Framework Convention on AI

José Jaime Villalobos,[*] and Matthijs M. Maas,[†] [1]

Abstract

Solving many of the challenges presented by artificial intelligence (AI) requires international coordination and cooperation. In response, the past years have seen multiple global initiatives to govern AI. However, very few proposals have discussed treaty models or design for AI governance and have therefore neglected the study of framework conventions–generally multilateral law-making treaties that establish a two-step regulatory process through which initially underspecified obligations and implementation mechanisms are subsequently specified via protocols. This chapter asks whether or how a Framework Convention on AI (FCAI) might serve as a regulatory tool for global AI governance, in contrast with the more traditional piecemeal approach based on individual treaties that govern isolated issues and have no subsequent regime. To answer these questions, the chapter first briefly sets out the recent context of global AI governance, and the governance gaps that remain to be filled. It then explores the elements, definition, and general role of framework conventions as an international regulatory instrument. On this basis, the chapter considers the structural trade-offs and challenges that an FCAI would face, before discussing key ways in which it could be designed to address these concerns. We argue that, while imperfect, an FCAI may be the most tractable and appropriate solution for the international governance of AI if it follows a hybrid model that combines a wide scope with specific obligations and implementation mechanisms concerning issues on which states already converge.

Keywords

Artificial intelligence, AI governance, international law, framework convention, treaty regimes, instrument choice, instrument design.

[1] For comments, we thank Duncan Cass-Beggs, Casey Mahoney, Carlos Ignacio Gutierrez, and Stephen Clare.

[*] Research Affiliate, Institute for Law & AI | Research Affiliate, Oxford Martin AI Governance Initiative. ORCID iD: 0000-0002-9015-6644. Corresponding author: Email: jose.villalobos@law-ai.org.

[†] Senior Research Fellow, Institute for Law & AI | Research Affiliate, Leverhulme Centre for the Future of Intelligence, University of Cambridge. ORCID iD: 0000-0002-6170-9393. Email: mmm71@cam.ac.uk.

Recent and ongoing progress in artificial intelligence (AI) technology has highlighted the technology's potential for increasingly significant global impacts.[2] These are due to the broad range of capabilities and applications AI systems can enable, which can threaten human rights,[3] disrupt international relations,[4] and gravely affect international peace and security.[5] In particular, there may be distinct and growing risks from advanced AI systems,[6] including increasingly general-purpose and capable foundation models,[7] dual-use foundation models, AI agents,[8] and other systems at the frontier that may create challenges stemming from their potential misuse, as well as increasingly hazardous emergent capabilities.[9]

[2] Nestor Maslej and others, 'The AI Index 2024 Annual Report' (AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, 2024); Gavin Leech and others, 'Ten Hard Problems in Artificial Intelligence We Must Get Right' (2024) arXiv <https://doi.org/10.48550/arXiv.2402.04464> accessed 7 July 2024.

[3] Lorna McGregor, Daragh Murray, and Vivian Ng, 'International Human Rights Law as a Framework for Algorithmic Accountability' (2019) 68(2) International & Comparative Law Quarterly 309; Kate Jones, 'AI Governance and Human Rights: Resetting the Relationship' (2023) Research Paper, International Law Programme, Chatham House <https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights> accessed 7 July 2024.

[4] Mary L Cummings and others, 'Artificial Intelligence and International Affairs: Disruption Anticipated' (2018) Chatham House Report <https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf>; Stephane J Baele and others, 'AI IR: Charting International Relations in the Age of Artificial Intelligence' (2024) 26(2) International Studies Review viae013 <https://doi.org/10.1093/isr/viae013> accessed 7 July 2024.

[5] Matthijs Maas, Kayla Lucero-Matteucci, and Di Cooke, 'Military Artificial Intelligence as a Contributor to Global Catastrophic Risk' in SJ Beard and others (eds) *The Era of Global Risk: An Introduction to Existential Risk Studies* (Open Book Publishers 2023).

[6] By 'advanced AI', we refer to a broad category, that comprises both general-purpose AI systems that exhibit strong, near-human performance across a wide range of tasks (e.g. Claude Opus), as well as narrow but paradigm-shifting AI systems which outperform humans on specific tasks (e.g. Alphafold). See Matthijs M Maas, *Architectures of Global AI Governance: From Technological Change to Human Choice* (OUP forthcoming 2025), ch 2. There are, of course, many other and adjacent terms.

[7] For various accounts and definitions of these models, see: Carlos I Gutierrez and others, 'A Proposal for a Definition of General Purpose Artificial Intelligence Systems' (2023) 2(3) Digital Society 36; Elizabeth Seger and others, 'Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives' (2023) Centre for the Governance of AI < https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models> accessed 7 July 2024.

[8] Peter Cihon, 'Chilling Autonomy: Policy Enforcement for Human Oversight of AI Agents' (*GitHub*, 2024) <https://blog.genlaw.org/pdfs/genlaw_icml2024/79.pdf> accessed 15 August 2024; Noam Kolt, 'Governing AI Agents' (2024) SSRN Scholarly Paper <https://papers.ssrn.com/abstract=4772956> accessed 7 July 2024.

[9] Department for Science, Innovation & Technology, *Frontier AI: Capabilities and Risks – Discussion Paper on the Need for Further Research into AI Risk* (2023) <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper> accessed 7 July 2024; Anwar Usman and others, 'Foundational Challenges in Assuring Alignment and Safety of Large Language Models' (2024) arXiv <https://arxiv.org/abs/2404.09932> accessed 7 July 2024; Department for Science, Innovation & Technology and AI Safety Institute, *International Scientific Report on the Safety of Advanced AI: Interim Report* (2024) DSIT Research Paper Series 2024/009, 2024; Yoshua Bengio and others, 'Managing Extreme AI Risks amid Rapid Progress' (2024) *Science* <https://doi.org/10.1126/science.adn0117> accessed 7 July 2024.

In response, intense attention has been directed at the question of how to regulate these technologies, both at domestic and international levels.[10] In particular, international governance solutions for AI are both urgent and undersupplied. Many of AI's challenges will gain from forms of international coordination and cooperation to be effectively addressed–and some issues may require it.[11] However, efforts to achieve effective global governance for AI face many challenges. For one, as is occurring in many other domains of international law, global AI governance efforts face a tense geopolitical context of heightened international contestation and conflict.[12] In the domain of AI, such efforts are likely to face the growing rhetoric of potential AI arms races.[13] Finally, cooperative efforts are beset by the ever-growing fragmentation of international law, expressed in the emerging regime complex of institutions active on AI issues.[14]

Where should global AI governance go next? There are a wide range of proposals, from calls to apply existing international law[15] to the adaptation of existing institutions for this role.[16] However, a particular focus in recent years has been on the potential creation of new international institutions, or even comprehensive legal frameworks–whether completely novel or inspired by the model of existing legal regimes.[17] These regimes are often designed to carry

[10] Jonas Tallberg and others, 'The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research' (2023) 25(3) International Studies Review viad040 <https://doi.org/10.1093/isr/viad040> accessed 7 July 2024; Michael Veale, Kira Matus, and Robert Gorwa, 'AI and Global Governance: Modalities, Rationales, Tensions' (2023) 19(1) Annual Review of Law and Social Science 19 <. https://doi.org/10.1146/annurev-lawsocsci-020223-040749> accessed 7 July 2024.

[11] Maas (n 5). For a general treatment of the conditions of international cooperation and coordination, see also Arthur A Stein, 'Coordination and Collaboration: Regimes in an Anarchic World' (1982) 36(2) International Organization 299.

[12] Karen J Alter, 'The Future of International Law' in Diana Ayton-Shenker (ed), *The New Global Agenda* Lahnham: Rowman & Littlefield 2018).

[13] Michael T Klare, 'AI Arms Race Gains Speed' (2019) 49(2) Arms Control Today 35; although for critiques of the framing, see also Haydn Belfield and Christian Ruhl, 'Why Policy Makers Should Beware Claims of New "Arms Races"' (*Bulletin of the Atomic Scientists*, 14 July 2022) <https://thebulletin.org/2022/07/why-policy-makers-should-beware-claims-of-new-arms-races/> accessed 7 July 2024.

[14] Peter Cihon, Matthijs M Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance' (2020) 11(5) Global Policy 545, 545-56; Huw Roberts and others, 'Global AI Governance: Barriers and Pathways Forward' (2024) SSRN Scholarly Paper <https://doi.org/10.2139/ssrn.4588040> accessed 8 July 2024. Emma Klein and Stewart Patrick, 'Envisioning a Global Regime Complex to Govern Artificial Intelligence' (2024) Carnegie Endowment for International Peace <https://carnegieendowment.org/research/2024/03/envisioning-a-global-regime-complex-to-govern-artificial-intelligence?lang=en> accessed 2 April 2024.

[15] See for instance: Martina Kunz and Seán Ó hÉigeartaigh, 'Artificial Intelligence and Robotization' in Robin Geiß and Nils Melzer (eds), *The Oxford Handbook of the International Law of Global Security* (OUP, 2021); Bryant W Smith, 'New Technologies and Old Treaties' (2020) 114 *AJIL Unbound* <https://doi.org/10.1017/aju.2020.28> accessed 8 July 2024, 152-57; Mark Chinen, *The International Governance of Artificial Intelligence* (Edward Elgar, 2023), ch. 6; Antonio Coco and Talita Dias, '"Handle with Care": Due Diligence Obligations in the Employment of AI Technologies' in Robin Geiß and Henning Lahmann (eds), *Research Handbook on Warfare and Artificial Intelligence* (Edward Elgar, 2024).

[16] Roberts and others (n 14); Rumtin Sepasspour, 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence' (2023) 79(5) Bulletin of the Atomic Scientists 304.

[17] Matthijs M Maas and José J Villalobos, 'International AI Institutions: A Literature Review of Models, Examples, and Proposals' (2023) Institute for Law & AI, AI Foundations Report <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4579773> accessed 8 July 2024; Lewis Ho and others, 'International Institutions for Advanced AI' (2023) arXiv <https://doi.org/10.48550/arXiv.2307.04699> accessed 8 July 2024.

out much-needed governance functions around AI, and to anchor international collaboration on managing the risks, as well as realising the benefits, of this technology.

Yet, in many cases, proposals of such institutions omit detailed analysis of questions of treaty design.[18] Instead, they trade largely on relatively light and unreflexive analogies,[19] which creates a risk of inappropriate institutional mimicry.[20] Then, the proposals that do suggest new treaties to govern AI often neglect treaty design and, with a few exceptions we will discuss below, the prospects of a framework convention to regulate AI.

In response, this chapter will explore the general role of framework conventions as an international regulatory instrument, to consider the prospects and design tradeoffs of a potential Framework Convention on Artificial Intelligence (FCAI). Specifically, we ask whether a framework convention would be the most optimal regulatory tool for the governance of problems presented by AI? What tradeoffs would it face, and how might it be designed to overcome these?

To answer these questions, this chapter is divided into four sections. In Section 1, we briefly describe recent developments in global AI governance, proposals for the way forward, the potential governance gaps that remain, and which of these could potentially be filled by an FCAI. In Section 2, we review the essence of framework conventions as regulatory instruments, by examining definitions and taxonomies of different types of international treaties; identifying the components and elements of framework conventions, and briefly reviewing their use and track record in different domains of international law. Section 3 then turns to consider the structural trade-offs facing an FCAI. In response to those challenges, Section 4 evaluates key design elements of such a convention. On this basis, we conclude that an FCAI may be the most tractable solution for the challenges presented by AI. Nevertheless, it runs the risk of being an ineffective regulatory tool, like many framework conventions before it, if it does not adopt a hybrid model that combines a wide scope with strong obligations and implementation mechanisms concerning certain governance issues.

1. A Fragmented AI Governance Regime

Recent years have not been a quiet time for the global governance of AI. Since the mid- and late-2010s, an initial wave of hundreds of non-binding AI ethics principles gradually culminated in two landmark (albeit) non-binding recommendations, by the OECD and UNESCO.[21] Then, since 2022, the period of intense public attention for 'generative AI' gave rise to a new wave of soft-law initiatives and declarations, such as notably the AI Safety

[18] Maas and Villalobos (n 17) 44.

[19] Matthijs Maas, 'AI Is Like... A Literature Review of AI Metaphors and Why They Matter for Policy' (2023) Institute for Law & AI, AI Foundations Report <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4612468>.

[20] For a general analysis of the conditions under which states pursue mimicking dynamics, see Bernhard Reinsberg and Oliver Westerwinter, 'Institutional Overlap in Global Governance and the Design of Intergovernmental Organizations' (2023) 18(4) The Review of International Organizations 693.

[21] OECD, 'Recommendation of the Council on Artificial Intelligence' (2019) OECD Legal Instruments OECD/LEGAL/0449 <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> accessed 8 July 2024; UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2021) <https://unesdoc.unesco.org/ark:/48223/pf0000381137> accessed 8 July 2024.

Summit series–with its *Bletchley*- and *Seoul Declarations*–as well as the G7 Hiroshima Process *Guiding Principles* and *International Code of Conduct*, amongst others.[22] Finally, 2024 saw major AI resolutions at the UN General Assembly,[23] alongside the first binding regulatory initiatives on AI, in the form of the EU AI Act and the Council of Europe's *Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law* (CoE Framework Convention).[24] While promising, there are also many outstanding critiques of the adequacy of some of these instruments to the size of the challenge. For instance, the EU AI Act, for all its benefits, has been criticised for incorporating potential oversight loopholes,[25] and non-binding club initiatives have been critiqued as potentially too weak or underinclusive.[26]

Consequently, proposals for a new international regime on AI have proliferated in recent years. In previous work, we reviewed and evaluated a range of proposals for new international AI institutions, identifying seven distinct models that have been offered by scholars and practitioners.[27] In most cases, proposals for new institutions on AI omitted a detailed analysis of treaty design questions.[28] Therefore, we highlighted that a direction for further research could be a study of multilateral AI treaty options, which is what we undertake in this chapter.

To be sure, there are many treaty proposals to govern AI. This includes various proposals to establish treaty restrictions on autonomous weapons,[29] such as calls for a 'Digital Geneva Convention',[30] or internationally codified bans on lethal autonomous weapons systems (LAWS), inspired by global bans on blinding lasers and anti-personnel mines.[31] There are also

[22] 'Hiroshima Process International Guiding Principles for Advanced AI System' (2023) <https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-guiding-principles-advanced-ai-system> accessed 8 July 2024; 'Hiroshima Process International Code of Conduct for Advanced AI Systems' (2023) <https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems> accessed 8 July 2024.

[23] UN Doc A/Res/L.49 (2024) [Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development]; UN Doc A/Res/78/311 (2024) [Enhancing international cooperation on capacity-building of artificial intelligence].

[24] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (EU AI Act) [2024] OJ L2024/1689; Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, opened for signature 5 September 2024, CETS 225.

[25] Sandra Wachter, 'Limitations and Loopholes in the E.U. AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26(3) Yale Journal of Law & Technology 671.

[26] Inga Ulnicane, 'Governance Fix? Power and Politics in Controversies about Governing Generative AI' (2024) Policy and Society puae022 <https://doi.org/10.1093/polsoc/puae022> accessed 9 July 2024.

[27] Maas and Villalobos (n 17) 44.

[28] ibid.

[29] Bonnie Docherty, 'The Need for and Elements of a New Treaty on Fully Autonomous Weapons' (*Human Rights Watch*, 1 June 2020) <https://www.hrw.org/news/2020/06/01/need-and-elements-new-treaty-fully-autonomous-weapons> accessed 14 July 2024.

[30] Brad Smith, 'The Need for a Digital Geneva Convention' (Microsoft on the Issues, 14 February 2017) <https://blogs.microsoft.com/on-the-issues/2017/02/14/need-digital-geneva-convention/> accessed 17 April 2019.

[31] Human Rights Watch, 'Precedent for Preemption: The Ban on Blinding Lasers as a Model for a Killer Robots Prohibition: Memorandum to Convention on Conventional Weapons Delegates' (*Human Rights Watch*, 8

proposals for international conventions to mitigate extreme risks from AI technology. Some of these address many technological risks in concert.[32] Others include a broader remit, such as proposals to create a *Universal Convention on Artificial Intelligence for Humanity* to uphold human values such as dignity, privacy and freedom.[33] Other treaty suggestions focus more specifically on regulating significant risks from advanced AI in particular, whether in the form of an 'AI development convention' that would set down 'strict safety rules for certain kinds of AI development',[34] restrictions or global moratoria on particular types of systems,[35] or treaties to preclude the use of AI systems in interstate warfare.[36]

Nonetheless, many of these recommendations derive from an underspecified understanding of the range of regulatory tools available–and the full space of both instrument choice and instrument design within international law. In particular, they reflect a focus on singular isolated treaties on AI issues, rather than an overarching legal framework. Given this, with a few exceptions that are discussed below, there has not yet been much attention for the potential role of a distinct regulatory instrument for AI: the framework convention.[37] As discussed in more detail below, this is a type of treaty that establishes underspecified obligations and implementation mechanisms that are then specified through subsequent protocols.

Actions and commentary on a framework convention for AI governance have only recently begun to mature. Most debates in this space have, understandably, focused on the

November 2015) <https://www.hrw.org/news/2015/11/08/precedent-preemption-ban-blinding-lasers-model-killer-robots-prohibition> accessed 28 April 2017.

[32] Grant Wilson, 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law' (2013) 31 Va. Envtl. LJ 307; Guglielmo Verdirame, 'For China, a Legal Reckoning Is Coming' (*UnHerd*, 20 April 2020) <https://unherd.com/2020/04/for-china-a-legal-reckoning-is-coming/> accessed 14 July 2024.

[33] Maral Niazi, 'Universal Convention on Artificial Intelligence for Humanity' (2024) Centre for International Governance Innovation <https://www.cigionline.org/publications/universal-convention-on-artificial-intelligence-for-humanity/> accessed 23 July 2024.

[34] Daniel Dewey, 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff' in Vincent C Müller (ed), *Risks of Artificial Intelligence* (Chapman and Hall/CRC 2015), 7-8.

[35] For such proposals, on various grounds, see Thomas Metzinger, 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology' (2021) 19 Journal of Artificial Intelligence and Consciousness 1; Jason Hausenloy, Andrea Miotti, and Claire Dennis, 'Multinational AGI Consortium (MAGIC): A Proposal for International Coordination on AI' (2023) arXiv <https://doi.org/10.48550/arXiv.2310.09217> accessed 14 July 2024.; Anthony Aguirre, 'Close the Gates to an Inhuman Future: How and Why We Should Choose to Not Develop Superhuman General-Purpose Artificial Intelligence' (2023) SSRN Scholarly Paper <https://papers.ssrn.com/abstract=4608505> accessed 14 July 2024.

[36] Elias G Carayannis and John Draper, 'Optimising Peace through a Universal Global Peace Treaty to Constrain the Risk of War from a Militarised Artificial Superintelligence' (2022) 11 AI & SOCIETY 2679 (discussing a 'Universal Global Peace Treaty', supported by a separate Cyberweapons and AI Convention).

[37] Of course, some scholars have proposed overarching 'frameworks' for global cooperation on AI. See for instance Pekka Ala-Pietilä and Nathalie A Smuha, 'A Framework for Global Cooperation on Artificial Intelligence and Its Governance' in Bertrand Braunschweig and Malik Ghallab (eds), *Reflections on Artificial Intelligence for Humanity* (Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021); Matthijs M Maas, 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks' (PhD thesis, University of Copenhagen 2020). However, these are usually analytical frameworks for organising the substantive areas and guiding principles involved in global AI governance, rather than specific proposals for the use of a framework convention instrument per se.

developments in instruments such as the CoE Framework Convention.[38] Yet, this instrument, while significant, hardly exhausts the potential scope, and potential role, of framework conventions in AI governance, and as such will not be the main focus of this chapter. This is both for substantive and analytical reasons. Substantively, while this convention is a significant achievement, and may well shape the global governance landscape for AI in coming years, it also presents significant drawbacks and hurdles. For instance, it focuses less on certain types of risks, especially those arising from more advanced models, and introduces obligations that are mostly redundant with those based on well-established international rules. It also makes exceptions for private actors, national security, pre-deployment activities, and national defence, which leaves significant regulatory gaps. It may also be less adaptive over time. Analytically, and more relevantly to our purposes, the CoE Convention is a plurilateral instrument–that is, one where membership to the treaty is 'closed' to a certain group of states unless States Parties give their unanimous consent to other states joining (see also Section 2).[39] Even without this hurdle, many states may choose not to sign the convention given their relatively limited involvement in its negotiation.[40] Thus, it is important to note that the hypothetical FCAI we will discuss throughout this paper would constitute a separate treaty and distinguish itself by aiming to be multilateral and by ideally neutralising many of the downsides that framework conventions, including the CoE Framework Convention, usually present.

Moreover, in a recent discussion paper, Cass-Beggs and others have envisioned a *Framework Convention on Global AI Challenges* aimed at achieving the three-fold purpose of realising and sharing the benefits of AI globally, mitigating severe global risks posed by AI, and ensuring legitimate and effective decisions can be made over the future of AI.[41] Our chapter aims to complement this work by analysing some aspects of treaty design in more detail, thus furthering a grounded evaluation of whether or when framework conventions are an appropriate regulatory tool–and if so, how to craft and design such an FCAI.

[38] Francesco P Levantino and Federica Paolucci, 'Advancing the Protection of Fundamental Rights Through AI Regulation: How the EU and the Council of Europe Are Shaping the Future' in Philip Czech and others (eds), *European Yearbook on Human Rights 2024* (Brill. Intersentia, 2024); Hannah van Kolfschooten and Carmel Shachar, 'The Council of Europe's AI Convention (2023–2024): Promises and Pitfalls for Health Protection' (2023) 138 Health Policy 104935; Marten Breuer, 'The Council of Europe as an AI Standard Setter' (*Verfassungsblog*, 4 April 2022) <https://verfassungsblog.de/the-council-of-europe-as-an-ai-standard-setter/> accessed 14 July 2024.

[39] Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (n 24) arts 30.1, 31.1.

[40] For instance, this was a challenge that initially also beset the 2001 Budapest Cybercrime Convention, equally negotiated by the CoE. While this did see some expansion to a number of non-CoE states, major states such as Brazil (until 2022) and India (to the present) declined to sign, in part because they had not been involved in negotiations. See Alexander Seger, 'India and the Budapest Convention: Why Not?' (orfonline.org, 10 August 2016) <https://www.orfonline.org/expert-speak/india-and-the-budapest-convention-why-not> accessed 14 July 2024, 7.

[41] Duncan Cass-Beggs and others, 'Framework Convention on Global AI Challenges: Accelerating International Cooperation to Ensure Beneficial, Safe and Inclusive AI' (2024) CIGI Discussion Paper, Centre for International Governance Innovation <https://www.cigionline.org/publications/framework-convention-on-global-ai-challenges/> accessed 25 June 2024.

2. The Essence and Elements of Framework Conventions

As we have established, AI poses many challenges that require international coordination and perhaps even binding commitments between states. While nations can affect each others' behaviours through non-binding legal norms,[42] or even through non-legal strategies such as shaming, rewarding, and socialisation,[43] binding obligations, as a matter of doctrine, come from three sources within international law: international treaties, customary international law, or general principles of law.[44] While they often receive less attention, the role of rules of customary international law and of general principles of law in the international governance of advanced AI should not be underestimated.[45] Nevertheless, for the purposes of this chapter, we will focus on the study of international treaties–and framework conventions in particular–as enablers of said governance.

International treaties–also referred to as international conventions, agreements, charters, pacts, or covenants, among other terms–are the oldest and 'the most important source of obligation in international law.'[46] The Vienna Convention on the Law of Treaties (VCLT) defines them as 'an international agreement concluded between states in written form and governed by international law, whether embodied in a single instrument or in two or more related instruments and whatever its particular designation.'[47] In other words, when in force, international treaties govern the conduct of states that have voluntarily acceded to become their parties.[48] On some occasions, they also set the constitutional mandate for international organisations.[49]

[42] Dinah Shelton (ed), *Commitment and Compliance: The Role of Non-Binding Norms in the International Legal System* (OUP 2000).

[43] Anne van Aaken and Betül Simsek, 'Rewarding in International Law' (2021) 115(2) American Journal of International Law 195; Robert Howse and Ruti Teitel, 'Beyond Compliance: Rethinking Why International Law Really Matters' (2010) 1(2) Global Policy; Timothy Meyer, 'How Compliance Understates Effectiveness' (2014) 108 AJIL Unbound; see generally Oona Hathaway, 'Between Power and Principle: An Integrated Theory of International Law' (2005) 72(2) University of Chicago Law Review; Ryan Goodman and Derek Jinks, *Socializing States: Promoting Human Rights Through International Law* (OUP 2013). Alex Geisinger and Michael A Stein, 'A Theory of Expressive International Law' (2006) 60(1) Vanderbilt Law Review 77.

[44] Statute of the International Court of Justice, art 38(1)(a.-c.); James Crawford, *Brownlie's Principles of International Law* (9th ed, Oxford University Press, 2019), 21-35.

[45] Among others, the principle of distinction and the no-harm principle. See Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law* (vol. 1, Cambridge University Press, 2012), rule 1; Pierre-Marie Dupuy and Jorge E. Viñuales, *International Environmental Law* (Cambridge University Press, 2015), 55–56.

[46] Crawford (n 44) 28; Malcolm N Shaw, *International Law* (8th ed, CUP 2017), 10; Randall Lessafer, 'Treaties within the History of International Law' in Michael J. Bowman and Dino Kritsiotis (eds), *Conceptual and Contextual Perspectives on the Modern Law of Treaties* (CUP 2018).

[47] Vienna Convention on the Law of Treaties, opened for signature 23 May 1969, 1155 U.N.T.S. 331, entered into force 27 January 1980, art 2(1)(a). See also Duncan B Hollis (ed), *The Oxford Guide to Treaties* (2nd ed, OUP 2020), 5, 19.

[48] Crawford (n 44) 28; Hollis (n 47) 38; see also International Law Commission, 'Draft Articles on the Law of Treaties with commentaries' (1966) Yearbook of the International Law Commission, vol. II, 188-189.

[49] See, for example, Constitution of the International Labour Organisation, 1 April 1919, Part XIII of the Treaty of Versailles, entered into force 28 June 1919; Constitution of the World Health Organization, opened for signature 22 July 1946, 14 U.N.T.S. 185, entered into force 7 April 1948.

International treaties come in many shapes, and many typologies of these sources of international law have been attempted.[50] In one meta-study of these typologies, Brölmann organises them under three wide categories: form, normative effect, and content.[51] Regarding their *form*, treaties can be bilateral, if they only have two states parties; multilateral, if they have more than two states parties and, in some cases, tend towards universality (most or all states being parties); or plurilateral, if there are multiple parties to a treaty, but these are divided into 'sides' or the treaty's membership is 'closed' to a certain group of states; among other possible classifications.[52]

In terms of *normative effect*, one typology that must be highlighted in the context of this chapter is that of normative completeness at the international level. On one hand, some treaties are are designed to be comprehensive, in that they aim to establish rules that can, in and of themselves, regulate a subject matter. These are contrasted with 'framework treaties' or 'continuing treaties', which are meant to be the starting point of a broader 'treaty regime', composed of a first general agreement setting the base for policy making and normative development in regards to a particular governance issue, followed by one or more subsequent agreements (normally referred to as 'protocols').[53] Another category of treaties worth mentioning in the context of normative effects, and regulatory function in particular, is that of *law-making treaties*; that is, conventions that create obligations with an inherent juridical force that is independent from implementation by other parties. These stand in contrast to *interdependent treaties*, by which the implementation of one party is dependent or conditioned by the performance of the other parties.[54]

Finally, treaties can be categorised according to their *content*. For instance, all instruments pertaining to the same area of law (e.g., international humanitarian law or international human rights law) could be classed together as one type of treaty.[55]

Where do framework conventions sit within these typologies and how can they be defined? To be clear, the VCLT does not define or establish separate rules concerning framework conventions, as it was adopted in 1969, years before framework conventions were established as a regulatory technique of public international law.[56] As mentioned above, Brölmann refers to framework conventions as a category of treaty that tends towards normative *in*completeness at first but then aims to gradually achieve normative completeness through subsequent protocols.[57] Indeed, this two-step regulatory process through which negotiators 'delegate questions that are relevant for achieving the agreement's objectives to additional regulation' is the main distinguishing factor of the 'framework convention and protocol

[50] See, for example OECD, *Compendium of International Organisations' Practices: Working Towards More Effective International Instruments* (OECD Publishing 2021).

[51] Catherine Brölmann, 'Typologies and the 'Essential Juridical Character' of Treaties' in Bowman and Kritsiotis (n 46).

[52] ibid 85-86.

[53] ibid 89-90.

[54] ibid 92-93.

[55] ibid 97-98.

[56] Nele Matz-Lück, 'Framework Conventions as Regulatory Tools' (2009) 1 Goettingen Journal of International Law 3, 451.

[57] Brölmann (n 51) 89.

approach'. In contrast, under the more traditional 'piecemeal approach', states only regulate an isolated aspect of a larger problem, as it stands at a given time of negotiation, and do not establish mechanisms for subsequent agreements.[58]

Nonetheless, a more holistic definition of framework conventions should locate them within other categories of treaties as well. For example, it is important to note that framework conventions are law-making, as they create obligations that states must fulfil regardless of other states' actions. Moreover, because they lack specificity and are therefore less politically controversial to create and enter into, framework conventions are also frequently multilateral, with an ambition to be universal.[59] Yet, most protocols to framework conventions are 'closed', in that only States Parties to the framework convention can subsequently become parties to its protocol(s).[60] Thus, we can define framework conventions as: *generally multilateral law-making treaties that establish a two-step regulatory process through which initially underspecified obligations and implementation mechanisms are subsequently specified via protocols.*

To establish the platform needed to develop a treaty regime through this two-step regulatory process, a framework convention normally contains at least three types of clauses. The first type of clause sets the *objectives* for both the framework convention and the treaty regime as a whole.[61] The second type of clause establishes *broad commitments* for the States Parties. These are normally specified through protocols, but are nonetheless binding in their vague form upon the framework convention's entry into force.[62] A number of parent conventions stipulate general obligations to, for instance, take 'measures', or 'steps', or 'develop policies and strategies', individually or in cooperation with other states, to meet the treaty's objectives.[63] Third, to ensure the continuation of the treaty regime that is being created, framework conventions usually have provisions to create *organs* that coordinate States Parties' actions towards enforcement and the negotiation of protocols, including a conference of the parties (COP) or meeting of the parties (MOP) and a secretariat to handle administrative

[58] Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 447.

[59] However, note that there are several examples of plurilateral framework conventions, due to political motives (e.g., the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law) or geographical reasons (e.g., the Framework Convention on the Protection and Sustainable Development of the Carpathians or the Agreement on the Nile River Basin Cooperative Framework), among other justifications.

[60] Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 441, 451. However, the implementation agreements of some framework conventions differ from protocols in that they are open to the membership of parties that are not parties to the parent convention. See Nele Matz-Lück, 'Framework Agreements' (*Max Planck Encyclopedias of International Law*, February 2011) <opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e703> accessed 8 August 2023, para 5.

[61] ibid 446.

[62] Nele Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 446.

[63] See, for example, Vienna Convention for the Protection of the Ozone Layer, Vienna, 22 Mar. 1985, 1513 U.N.T.S. 293, T.I.A.S. No. 11,087, 26 I.L.M. 1516 (1987), entered into force 22 Sept. 1988, art 2; International Covenant on Economic, Social and Cultural Rights, New York, 16 Dec. 1966, 993 U.N.T.S. 3, entered into force 3 Jan. 1976 (ICESCR), art 2.1; Convention on Long-Range Transboundary Pollution, opened for signature 14 November 1979, 1302 U.N.T.S. 217, entered into force 16 March 1983, art 3.

matters.[64] Two of the better known examples of these organs are the COPs pertaining to the United Nations Framework Convention on Climate Change (UNFCCC) and the CBD.[65]

Of course, taken alone, these are not design features that are exclusive to framework conventions. Moreover, some framework conventions follow a hybrid approach by establishing general objectives and commitments on various aspects regarding the regulated subject matter, while at the same time introducing specific clauses regarding a particular issue within that subject matter. Thus, issues on which States Parties have reached a consensus are narrowly regulated in the 'parent convention' and more controversial or uncertain questions are left for additional protocols.[66]

3. The Trade-offs of a Framework Convention on AI

Designing an international treaty as a framework convention can certainly facilitate the completion of an agreement in a politically contested context. Among other reasons, framework conventions have increasingly proven an apt avenue by which states can shift international governance towards the use of collective decision-making bodies–such as standing 'legislative bodies' like COPs–because the design of a framework convention can serve as a costly commitment device.[67] As noted by Meyer, at best, such mechanisms can help in 'limiting states' ability to choose which other states to negotiate with on an issue-by-issue basis and thereby ensuring states that they will have the right to participate in future negotiations.'[68]

Framework conventions also come with significant potential drawbacks that should be carefully considered and weighed against their advantages. Below, we will explore some of the trade-offs between framework conventions and an alternative piecemeal approach of negotiating one or several independent treaty agreements, containing more detailed norms, but with a narrower scope. Moreover, we will discuss what the trade-offs might be between those options in the unique context of an FCAI.

One benefit of an FCAI could lie in the way it might bridge currently emerging fissures in the global regime complex for AI governance. As briefly mentioned above, the geopolitical status quo of recent years has given rise to many factors that hamper most forms of new international agreements between states, especially a new, binding international treaty.[69] In most cases, this might mean that any substantive agreements on any given topic might only be

[64] Nele Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 444.

[65] United Nations Framework Convention on Climate Change, New York, 9 May 1992, 1771 U.N.T.S. 107, 31 I.L.M. 849 (1992), entered into force 21 Mar. 1994 [UNFCCC], art 7; Convention on Biological Diversity, Rio de Janeiro, 5 June 1992, 1760 U.N.T.S. 79, entered into force 29 December 1993, art 23.

[66] Nele Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 449 (giving the example of the Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and their Disposal, which established detailed obligations regarding the transportation of waste but left the matter of liability for a subsequent protocol).

[67] Timothy Meyer, 'Collective Decision-Making in International Governance' (2014) 108 AJIL Unbound 30.

[68] ibid 30-1.

[69] See also Curtis Bradley, Jack Landman Goldsmith, and Oona A. Hathaway, 'The Rise of Nonbinding International Agreements: An Empirical, Comparative, and Normative Analysis' (2023) 90(5) The University of Chicago Law Review 1281.

possible, if at all, amongst smaller groups of like-minded states. In this situation, effective action might take the form of either minilateral action within informal club fora,[70] or at best take the form of plurilateral treaties negotiated by a few allied states. In the context of AI, if these current political and governance trends in the AI regime complex continue,[71] this would suggest the feasibility of one or two distinct substantive treaties. For instance, one regime might emerge amongst the United States and its allies (e.g., the United Kingdom, the European Union, South Korea, Japan) that focuses on issues around AI model safety, but where obligations are tailored so as to not impose measures or limitations that are seen as overly restrictive, as to maintain their advantage in the field.[72] Simultaneously, and in reaction, another treaty regime might be concluded amongst a coalition of Global South countries that are or feel excluded by such an arrangement.[73] Such a regime might be led, coordinated or nurtured by China, in particular under its Global AI Governance Initiative,[74] and instead emphasise strong benefit-sharing obligations and mandating strict mitigation measures concerning AI risk. Such an outcome could, for instance, echo previous instances of governance fragmentation,[75] competitive regime creation,[76] or forum shopping.[77] Such an outcome might be both normatively and functionally undesirable due to its role in splitting international law for AI, leading to potential norm interface conflicts, as well as cementing a more exclusive hierarchy.

[70] Jean-Frédéric Morin and others, 'How Informality Can Address Emerging Issues: Making the Most of the G7' (2019) 10(2) Global Policy 267.

[71] This is indeed reflected in the relative areas of focus of the recent UNGA Resolutions led by the US and China. See UN Doc A/78/L.49 (2024) [Seizing the Opportunities of Safe, Secure, and Trustworthy Artificial Intelligence Systems for Sustainable Development]; UN Doc A/78/L.86 (2024) [Enhancing International Cooperation on Capacity-Building of Artificial Intelligence].

[72] See, for example, Andrew Imbrie and others, 'Agile Alliances: How the United States and Its Allies Can Deliver a Democratic Way on AI' (2020) Center for Security and Emerging Technology <https://cset.georgetown.edu/wp-content/uploads/CSET-Agile-Alliances.pdf> accessed 2 August 2021; Leopold Aschenbrenner, *Situational Awareness: The Decade Ahead* <situational-awareness.ai/> accessed 5 July 2024, 137-38.

[73] Sumaya N Adan, 'The Case for Including the Global South in AI Governance Discussions' (*GovAI Blog*, 20 October 2023) <https://www.governance.ai/post/the-case-for-including-the-global-south-in-ai-governance-conversations> accessed 5 July 2024; Cecil Abungu, Michelle Malonza, and Sumaya N Adan, 'Can Apparent Bystanders Distinctively Shape an Outcome? Global South Countries and Global Catastrophic Risk-Focused Governance of Artificial Intelligence' (2023) arXiv <https://doi.org/10.48550/arXiv.2312.04616> accessed 5 July 2024.

[74] Chinese Ministry of Foreign Affairs, 'Global AI Governance Initiative' (*Chinese Ministry of Foreign Affairs*, 20 October 2023) <https://www.fmprc.gov.cn/eng/xw/zyxw/202405/t20240530_11332389.html> accessed 8 May 2024.

[75] See Frank Biermann and others, 'The Fragmentation of Global Governance Architectures: A Framework for Analysis' (2009) 9(4) Global Environmental Politics 14.

[76] Julia C Morse, and Robert O Keohane, 'Contested Multilateralism' (2014) 9(4) The Review of International Organizations 385, 398-406 (discussing how coalitions of states and non-state actors repeatedly challenged the World Health Organization (WHO), leading to a proliferation of regime complexes in global health). But for other strategies, see also Mette Eilstrup-Sangiovanni and Daniel Verdier, 'To Reform or to Replace? Institutional Succession in International Organizations' (European University Institute, 2021) <https://cadmus.eui.eu//handle/1814/69862> accessed 15 July 2024 (discussing institutional replacement as a common alternative to either institutional reform, 'regime-shifting' or 'competitive regime-creation').

[77] See Douglas W Gates, 'International Law Adrift: Forum Shopping, Forum Rejection, and the Future of Maritime Dispute Resolution' (2017) 18(1) Chicago Journal of International Law 287.

An FCAI might be able to bridge this type of political fissure, as other framework conventions have done in the past.[78] It could establish initially vague, but shared, obligations concerning key AI governance issues that are of interest to both sides. For example, this may be done by incorporating initially underspecified obligations with regards to specific issue areas–such as the mitigation of risks from advanced AI systems–[79]or in prescribing general responses to many imminent and currently pervasive impacts of AI systems on society, in domains such as disparate impacts or DeepFake misuse.[80] An FCAI could also set down broad commitments on benefit-sharing. More specific regulation of several specific questions could then be achieved through subsequent protocols (e.g., a protocol on AI safety and model evaluation).[81] Moreover, a framework convention might decrease the likelihood of a fragmented international legal regime on AI–which is, increasingly, a general-purpose technology dependent on a concentrated development chain–if such an FCAI encapsulates a large set of AI governance issues within a single treaty regime.[82]

Thus, reaching an agreement of such a wide scope is certainly tempting. While there are also risks of creating non-functional or underpowered 'empty institutions',[83] any type of agreement might arguably be better than no agreement at all, if that would leave large swathes of key AI governance issues largely unregulated.[84] Even if a vague treaty is likely to leave significant gaps on key governance issues, those might be slowly filled through follow-up institutions, such as COPs or networks of national contact points, as States Parties further refine their commitments or establish implementation mechanisms that would not have existed otherwise. General obligations in framework conventions tend to set a minimum threshold for compliance, but no explicit ceiling.[85] This would allow ambitious governments to pass

[78] See, for example, Stephanie Carvin, ‘Conventional Thinking? The 1980 Convention on Certain Conventional Weapons and the Politics of Legal Restraints on Weapons during the Cold War’ (2017) 19 Journal of Cold War Studies 1, 38-69.

[79] James N Baker, ‘International Law and Advanced AI: Exploring the Levers for “Hard” Control’ (*Institute for Law & AI*, 20 July 2024) <https://law-ai.org/international-law-and-advanced-ai/> accessed 17 August 2024; Stephan Llerena, ‘Global Governance of High-Risk Artificial Intelligence’ (2023) University of Chicago Existential Risk Laboratory <https://xrisk.uchicago.edu/files/2024/02/FINALLlerenaXLabSRF23-GlobalGovernanceOfHighRiskAI-0334e1dd25e37cae.pdf> accessed 15 December 2023.

[80] For a compendium on these and other risks from AI, see Peter Slattery and others, ‘The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence’ (2024) Massachusetts Institute of Technology AI Risk Repository <https://airisk.mit.edu/> accessed 21 August 2024.

[81] See, for example, the proposed Protocol on Global Public Safety and Security Risks from AI by Cass-Begs and others (n 41) 14-18.

[82] See Matz-Lück, ‘Framework Conventions as Regulatory Tools’ (n 56) 447. On the other hand, there is also the risk that if different states ratify some but not all additional protocols to a framework convention, subject to differing reservations, this undercuts the cohesion of the overall regime. See Rebecca Crootof, ‘Jurisprudential Space Junk: Treaties and New Technologies’ in Chiara Giorgetti and Natalie Klein (eds), *Resolving Conflicts in the Law* (Brill 2019), 120-121.

[83] Radoslav S Dimitrov, ‘Empty Institutions in Global Environmental Politics’ (2020) 22(3) International Studies Review 626.

[84] Radoslav S Dimitrov and others, ‘International Nonregimes: A Research Agenda’ (2007) 9(2) International Studies Review 230.

[85] See, for example, WHO Framework Convention on Tobacco Control, opened for signature 21 May 2003, 2302 U.N.T.S. 166, entered into force 27 February 2005, art 6 (indicating that ‘measures [to reduce the demand of tobacco] may include’ tax policies, price policies, and the prohibition or restriction of tax- or duty-free tobacco, but allowing states to take other measures); International Covenant on Economic, Social and Cultural Rights (n 63) art 2.1 (setting an obligation to progressively realise the rights in the Covenant ‘by all appropriate means’).

domestic legislation and regulations that go well beyond that minimum threshold. Moreover, a delayed two-stage process might allow for the eventual creation of more scientifically informed, politically robust, and substantive protocols.[86]

Indeed, the flexibility associated with a two-stage process might be ideal in the case of emerging technologies like AI, which are subject to the Collingridge dilemma.[87] Given that there are many uncertainties surrounding the extent of AI systems' capabilities, and the most adequate responses to the different risks those capabilities present, it might be more favourable to delay specific regulation to subsequent protocols, and then introduce more effective governance mechanisms into those instruments.[88] By contrast, an individual treaty agreement born out of pressure and rushed compromises could lead to a suboptimal or ineffective treaty regime.[89] Indeed, an attempt at producing a single specific treaty under a piecemeal approach may lock in the international legal regime on AI for decades, given the recorded difficulty of effectively amending multilateral international treaties once set down.[90] Thus, preserving the option value of an adequate regulatory framework is one strong argument in favour of a framework convention.

However, the watered down obligations and implementation mechanisms that are likely to come hand-in-hand with the extended scope of a framework convention and protocol approach also have important downsides. Just as some governments might seek to surpass the minimum threshold obligations established by a framework convention, others might read those obligations as a ceiling or implement only the measures required to meet that threshold, delaying actions that are urgently needed to curb global risks.[91]

An FCAI would not be immune to these challenges. For one, AI presents global risks that are potentially even more immediate or harder to repair than those presented by climate change. This means it is questionable whether the international community can afford to establish broad, unspecific obligations from the start. Moreover, as has been the case with many other framework conventions,[92] additional protocols to the FCAI may end up taking years to be negotiated and enter into force. As argued by Matz-Lück, the need to reach a framework convention is already an indication of a political stalemate, which implies that the negotiation of any protocol with specific and enforceable obligations will itself be a long and difficult

[86] Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 445.

[87] David Collingridge, *The Social Control of Technology* (Palgrave Macmillan 1981).

[88] See Jonas Schuett and others, 'From Principles to Rules: A Regulatory Approach for Frontier AI' in *The Oxford Handbook on the Foundations and Regulation of AI* (OUP forthcoming).

[89] Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 451.

[90] Brian Israel, 'Treaty Stasis' (2014) 108 AJIL Unbound 63; Crootof (n 82).

[91] See, for example, See Isak Stoddard and others, 'Three Decades of Climate Mitigation: Why Haven't We Bent the Global Emissions Curve?' (2021) 46 Annual Review of Environment and Resources 653, 659-661 (discussing how the UNFCCC's lack of concrete obligations and implementation mechanisms could be held at least partially responsible for states' delayed reduction of greenhouse gas emissions).

[92] For example, the Kyoto Protocol and the Paris Agreement entered into force in 2005 and in 2016 respectively–eleven and twenty-two years after the UNFCCC entered into force. The Cartagena Protocol and the Nagoya Protocol entered into force in 2003 and in 2014–ten and twenty-one years after the Convention on Biological Diversity entered into force. However, note that in the case of the Vienna Convention for the Protection of the Ozone Layer and its Montreal Protocol, only four months elapsed between each agreement's entry into force (and less than three years between their dates of conclusion). For access to this and other data related to times of entry into force of treaties, see generally United Nations, 'United Nations Treaty Series Online' <https://treaties.un.org>.

task.[93] Such a scenario might also be the reason most framework conventions do not include a (strong) implementation mechanism,[94] unlike the mechanisms found in some treaties that follow a piecemeal approach and tackle other global risks like weapons of mass destruction, which are generally embedded with much more specificity in the agreement and tend to be more robust.[95]

4. Designing a Framework Convention on Artificial Intelligence

Understanding the relative trade-offs between the framework convention and protocol approach and the piecemeal approach provides a clearer picture of what an ideal FCAI–which takes the best elements of the former approach and reduces its downsides as much as possible–might look like.

An FCAI should capitalise on what is perhaps the biggest advantage of framework conventions–their wide scope. By encompassing a large proportion of international governance issues related to AI, an FCAI can set a clear trajectory for a single general treaty regime on AI, rather than a fragmented one. For instance, the agreement could embed the legal obligations and institutional mechanisms to: prohibit certain unacceptable types or uses of AI; build scientific consensus on the technology's impacts, trajectories and risks; develop and harmonise standards related to risk management and safety; share or distribute the benefits of AI; carry out international joint research; put in place adequate emergency response measures; and monitor compliance; among others.[96]

While a framework convention might be necessary to bridge existing political gaps, there are certainly areas of AI governance which all or most states agree on. As per recent resolutions at the UN General Assembly, it seems clear that Member States share a commitment towards ensuring that AI systems are aligned with human rights and international law, that their benefits are enjoyed by all of humanity, and that they are 'safe, secure, and trustworthy'.[97] Furthermore, there seems to be a growing view among governments and

[93] Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 452-3.

[94] See, for example, Robert J Mathews, 'The 1980 Convention on Certain Conventional Weapons: A Useful Framework despite Earlier Disappointments' (2001) 83(844) International Review of the Red Cross 991, 997.

[95] See, for example, Treaty on the Non-Proliferation of Nuclear Weapons, Washington D.C., 1 July 1968, 729 U.N.T.S. 161, entered into force 5 Mar. 1970, art 3 (establishing the obligation to accept safeguards); United Nations, Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects (and Protocols) (As Amended on 21 December 2001), 1342 U.N.T.S. 137, 10 October 1980, art XII (giving the Conference of Parties the authority to take measures to ensure compliance with the treaty). For a general discussion of lessons for AI governance from other risks, see Seth Baum, 'Lessons for Artificial Intelligence from Other Global Risks' in Maurizio Tinnirello (ed), *The Global Politics of Artificial Intelligence* (CRC Press 2020).

[96] Maas and Villalobos (n 17); UN Advisory Body on Artificial Intelligence, 'Interim Report: Governing AI for Humanity' (2023) <https://www.un.org/sites/un2.un.org/files/un_ai_advisory_body_governing_ai_for_humanity_interim_report.pdf> accessed 22 December 2023, 15; Lewis Ho and others, 'International Institutions for Advanced AI' (2023) arXiv <https://arxiv.org/abs/2307.04699> accessed 1 August 2023, 6.

[97] UN Doc A/78/L.49 (n 23), paras 4, 6(a); UN Doc A/78/L.86 (n 23), paras 2, 7, 8.

technical experts that certain types or uses of AI should be strictly prohibited.[98] Thus, an FCAI could establish clear red lines regarding AI development, including, for instance, the prohibition to develop or deploy highly advanced AI systems that could undertake longer-range planning,[99] carry risks of being potentially misaligned,[100] or that could very easily and widely be misused. On such common-ground issues, states should move forward with specific obligations and implementation mechanisms similar to those that are typically found in a treaty that follows a piecemeal approach. As such, an FCAI would ideally be a hybrid convention, one that includes both vague commitments as well as detailed obligations, depending on the level of political will and technical certainty surrounding each governance issue.[101]

As seen above, one downside of framework conventions is that they tend to only make only vague references to implementation mechanisms, or omit them entirely, usually delegating them to protocols. Given that progress in AI development is still advancing at a very fast pace, and that the harms associated with this type of technology are only likely to increase as AI systems become more advanced and embedded in society, the obligations established by an FCAI should instead be implemented by states as soon as possible, and an existing or new international institution should be able to verify compliance immediately.[102] If a detailed implementation and verification system is not politically viable, an FCAI could at least introduce basic systems that can be developed further by a COP without necessarily requiring a protocol.

Other, more minor drawbacks of the framework convention and protocol approach could also be countered through better instrument design choices during the development of an FCAI. Firstly, while most framework conventions are ambiguous in terms of what the focus of its protocols should be, an FCAI could list a series of (prospective) protocols that states must negotiate as a matter of obligation, perhaps even within set time periods. This list should not be exhaustive, as other unforeseen problems or topics that require States Parties' coordination might of course emerge.[103] However, by setting out a concrete agenda and set of protocols for the agreement's COP to organise negotiations on, it is more plausible that meaningful protocols–which otherwise might take decades to be enacted–can instead be incorporated into the treaty regime with less delay.

Secondly, contrary to most framework conventions, an FCAI could introduce economic incentives for states to comply with their obligations. This link between legal obligations and economic incentives has been empirically demonstrated to increase the effectiveness of a treaty

[98] Akash Wasil and Tim Durgin, 'US-China Perspectives on Extreme AI Risks and Global Governance' (2024) SSRN Scholarly Paper <https://doi.org/10.2139/ssrn.4875436> accessed 22 August 2024; EU Artificial Intelligence Act (n 24), art 5; Future of Life Institute, 'Open Letter to the United Nations Convention on Certain Conventional Weapons' (*FLI*, 20 August 2017) <https://futureoflife.org/open-letter/autonomous-weapons-open-letter-2017/> accessed 18 March 2020.

[99] Michael K Cohen and others, 'Regulating Advanced Artificial Agents' (2024) 384(6691) Science 36.

[100] Bengio and others (n 9).

[101] Michael J Gilligan, 'Is There a Broader-Deeper Trade-off in International Multilateral Agreements?' (2024) 58(3) International Organization 459.

[102] For examples of the type of verification mechanisms that States could incorporate into an FCAI, see Akash R Wasil and others, 'Verification methods for international AI agreements' (2024) arXiv <https://arxiv.org/pdf/2408.16074> accessed 30 August 2024.

[103] Matz-Lück, 'Framework Conventions as Regulatory Tools' (n 56) 453.

at achieving its objectives.[104] For instance, an FCAI could condition access to other states' markets to compliance with international standards on due diligence or risk management.[105] It could also condition access to certain AI-related benefits on adherence to the treaty's obligations, similarly to the nuclear non-proliferation regime.[106] This could create a strong incentive for states to join an FCAI or even comply with its obligations even if they are not States Parties, as has been demonstrated by the so-called Brussels-, California-, or Beijing-effects of domestic or regional agreement with extraterritorial impacts.[107]

These instrument design choices would not guarantee the success of an FCAI. Nevertheless, they would go a long way towards addressing its drawbacks, and strengthen the case for pursuing this strategy instead of a piecemeal approach to the governance of AI.

Conclusion

International law is the realm of the imperfect–and of the necessary. An FCAI may not be the idealised legal solution to the challenges presented by AI. Nevertheless, it might be the only tractable avenue ahead, given the multiple geopolitical obstacles as well as given the uncertainty regarding the full extent of AI capabilities and optimal policy solutions. In this context, an FCAI would have the advantage of bringing together various groups under a single treaty regime, resulting in a unitary rather than a fragmented international governance regime on this critical technology. Additionally, such a framework convention and protocol approach provides the flexibility required for the governance of an emerging technology like AI. Nevertheless, there are significant downsides to a framework convention on AI that should be seriously considered and addressed. Among others, a framework convention might create an ineffective treaty regime that creates the illusion of comprehensive international regulation and drowns out more optimal regulatory solutions. Additionally, as seen with other framework conventions, some states might be willing to ratify a parent convention with vague commitments, but then refuse to become States Parties to subsequent protocols that actually specify those commitments through concrete obligations and implementation mechanisms.

However, there does not have to be a binary trade-off between either a framework convention and protocol approach, or a piecemeal approach. By adopting a hybrid design, an FCAI could bring together a wide range of states without compromising ambitious legal

[104] Steven J Hoffman and others, 'International Treaties Have Mostly Failed to Produce Their Intended Effects' (2022) 119(32) Proceedings of the National Academy of Sciences e2122854119 <https://doi.org/10.1073/pnas.2122854119> accessed 12 April 2023.

[105] See, for example, Robert Trager and others, 'International Governance of Civilian AI: A Jurisdictional Certification Approach' (2023) arXiv <https://doi.org/10.48550/arXiv.2308.15514> accessed 25 August 2023.

[106] Harry Law and Lewis Ho, 'Can a Dual Mandate Be a Model for the Global Governance of AI?' (2023) 5 Nature Reviews Physics 706, 706-7.

[107] Charlotte Siegmann and Markus Anderljung, 'The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market' (2022) Centre for the Governance of AI <https://www.governance.ai/research-paper/brussels-effect-ai> accessed 7 May 2023; Matthew S Erie and Thomas Streinz, 'The Beijing Effect: China's "Digital Silk Road" as Transnational Data Governance' (2021) 54 New York University Journal of International Law and Politics 1; Jens Frankenreiter, 'Cost-Based California Effects' (2022) 39 Yale Journal on Regulation 1098.

obligations in areas of AI governance where states do seem to agree, such as red lines or AI alignment. An FCAI could also introduce general implementation mechanisms–including verification ones–that can be applied immediately and be gradually fine-tuned by the treaty's COP, without having to delegate this task to a protocol. An FCAI could also include a (non-exhaustive) list of protocols to negotiate in the future, to bind states to negotiation processes, as well as strong economic incentives to participate in the treaty regime if they do not want to lose access to markets or AI-related benefits.

Of course, many more questions related to the international governance of AI need to be addressed before, and in order to, design an optimal legal framework. Regardless, states must start accepting the inevitability of a global agreement on what is perhaps the most crucial international coordination question of our time.