International AI Institutions／

A literature review of models, examples, and proposals

_AI Foundations Report 1

Matthijs Maas_
José Jaime Villalobos_

Legal Priorities Project

September 2023

International AI Institutions

A literature review of models, examples, and proposals

Legal Priorities Project – AI Foundations Report 1

September 2023

Matthijs M. Maas[1] **José Jaime Villalobos**[2 3]

Abstract: The question of how to ensure adequate international governance of artificial intelligence (AI) has come to the center of global attention. This literature review examines the range of institutional models that have been proposed as the basis for new international organizations focused on AI. It reviews and discusses these proposals for new international AI institutions, under a taxonomy of seven distinct institutional models that have been offered by scholars and practitioners. The models we include in this review are: (1) scientific consensus-building; (2) political consensus-building and norm-setting; (3) coordination of policy and regulation; (4) enforcement of standards or restrictions; (5) stabilization and emergency response; (6) international joint research; and (7) distribution of benefits or access.

For each model, we provide (i) a description of the model's functions and types; (ii) the most common examples of each model; (iii) some examples that are somewhat underexplored in the literature but that show promise; (iv) a review of proposals for the application of that model to the international regulation of AI; and (v) critiques of the model both generally and in its potential application to AI. In sum, we review more than thirty-two commonly invoked examples of these institutional models, twenty-three rarely-explored but promising alternate institutional examples, and forty-six proposals for new AI institutions. Finally, we sketch five directions for further research.

Cite as: Maas, Matthijs, and Villalobos, José Jaime. 'International AI Institutions: a literature review of models, examples, and proposals.' Legal Priorities Project, AI Foundations Report #1. (2023). https://www.legalpriorities.org/research/international-ai-institutions

[1] Senior Research Fellow, Legal Priorities Project | Research Affiliate, Centre for the Study of Existential Risk, University of Cambridge. ORCID iD: 0000-0002-6170-9393. Email: matthijs.maas@legalpriorities.org.

[2] International Law and Policy Fellow, Legal Priorities Project. ORCID iD: 0000-0002-9015-6644. Email: jose.villalobos@legalpriorities.org.

[3] For comments and input, we thank Matteo Pistillo, Cullen O'Keefe, Sumaya Nur Adan, and Harry Law. We also thank Alfredo Parra and Suzanne van Arsdale for support in preparing this report.

Executive Summary

This literature review examines a range of institutional models that have been proposed for the international governance of artificial intelligence (AI). The review specifically focuses on proposals that would involve the creation of new international institutions for AI. As such, it focuses on *seven* models for international AI institutions with distinct functions.

Part I consists of the literature review. For each model, we provide (i) a description of each model's functions and types; (ii) the most common examples of each model; (iii) some underexplored examples that are not (often) mentioned in the AI governance literature but that show promise; (iv) a review of proposals for the application of that model to the international regulation of AI; and (v) critiques of the model both generally and in its potential application to AI.

Part II briefly discusses some considerations for further research concerning the design of international institutions for AI, including the effectiveness of each model at accomplishing its aims; treaty-based regulatory frameworks; other institutional models not covered in this review; the compatibility of institutional functions; and institutional options to host a new international AI governance body.

Overall, the review covers seven models, as well as more than thirty-two common examples of those models, twenty-three additional examples, and forty-seven proposals of new AI institutions based on those models. Table 1 summarizes these findings.[4]

Table 1: Overview of institutional models, examples, and proposed institutions surveyed

Model	Common examples	Under-explored examples	Proposed AI institutions
1. Scientific consensus-building	• IPCC • IPBES • SAP	• CEP • WMO	• IPAI • Commission on Frontier AI • Intergovernmental Panel on Information Technology
2. Political consensus-building and norm-setting	• COPs (e.g. UNFCCC COP) • OECD • G20 • G7 • ISO • IEC • ITU • Various soft law	• Lysøen Declaration • Codex Alimentarius Commission • BRICS	• IAIO • Emerging Technology Coalition • IAAI • Data Governance Structure • Data Stewardship Organization • International Academy for AI Law and Regulation

[4] This table only contains a summary of (ii)–(iv) for each model. More details on the (i) functions and types of each model, and on (v) critiques of proposals for each model, can be found below.

	instruments		
3. Coordination of policy and regulation	• WTO • ICAO • IMO • IAEA • FATF • UNEP	• ILO • UNESCO • EMEP • World Bank • IMF • WSIS	• Advanced AI Governance Organisation • IAIO • EU AI Agency • GAIA • Generative AI global governance body • Coordinator and Catalyser of International AI Law
4. Enforcement of standards or restrictions	• IAEA (Department of Safeguards) • Nuclear Suppliers Group • Wassenaar Arrangement • Missile Technology Control Regime • Open Skies Consultative Commission • Atomic Development Authority	• OPCW • BWC Implementation Unit • IMO • CITES Secretariat	• UN AI control agency • Global watchdog agency • International Enforcement Agency • Emerging Technologies Treaty • IAIA (multiple) • UN Framework Convention on AI (UNFCAI) & Protocol on AI, supported by Intergovernmental Panel on AI, AI GLobal Authority, and supervisory body • Advanced AI Governance Organization • AIEA for Superintelligence • NPT+ • Multilateral AI governance initiative • International AI Safety Agency • Advanced AI chips registry • Code of conduct for state behavior • AI CBMs • Open Skies for AI • Bilateral US-China regime
5. Stabilization and emergency response	• FSB	• WHO • IAEA • UNDRR	• Geotechnology Stability Board
6. International joint research	• CERN • ITER • ISS • Human Genome Project • Atomic Development Authority (proposed)	• James Webb Telescope • LIGO	• AI Safety Project • Clearinghouse for research into AI • Benevolent AGI Treaty • Multilateral Artificial Intelligence Research Institute (MAIRI) • Neutral hub for AI research • UN AI Research Organization (UNAIRO) • CERN for AI • International supercomputing research facility • Joint international AI project • Multilateral AGI Consortium • European Artificial Intelligence megaproject
7. Distribution of benefits and access	• Gavi Vaccine Alliance • Global Fund to Fight AIDS, Tuberculosis and Malaria • IAEA (nuclear fuel bank)	• ABS Clearing-House • UN Climate Technology Centre and Network • UNIDO	• International Digital Democracy Initiative • Frontier AI Collaborative • Institution analogous to the IAEA • Fair and Equitable Benefit Sharing Model

Table of Contents

Executive Summary **3**

Introduction **6**

I. Review of Institutional Models **13**

Model 1: Scientific consensus-building 13

Model 2: Political consensus-building and norm-setting 17

Model 3: Coordination of policy and regulation 21

Model 4: Enforcement of standards or restrictions 24

Model 5: Stabilization and emergency response 32

Model 6: International joint research 33

Model 7: Distribution of benefits and access 38

II. Directions for Further Research **40**

Direction 1: Effectiveness of institutional models 40

Direction 2: Multilateral AI treaties without institutions 43

Direction 3: Additional institutional models not covered in detail in this review 44

Direction 4: Compatibility of institutional functions 45

Direction 5: Potential fora for an international AI organization 46

Conclusion **48**

Introduction

Recent and ongoing progress in artificial intelligence (AI) technology has highlighted that AI systems will have increasingly significant global impacts. In response, the past year has seen intense attention to the question of how to regulate these technologies, both at domestic and international levels. As part of this process, there have been renewed calls for the establishment of new international institutions to carry out much-needed governance functions and anchor international collaboration on managing the risks, as well as realizing the benefits, of this technology.

This literature review examines and categorizes a wide range of institutions that have been proposed to carry out the international governance of AI.[5] Before reviewing these models, however, it is important to situate proposals to establish a new international institution on AI within the broader landscape of approaches to the global governance of AI. Not all approaches to AI governance focus on the creation of new institutions. Rather, the institutional approach is only one of several different approaches to international AI governance–each of them concentrating on different governance challenges posed by AI, and each of them providing different solutions.[6] These approaches include:

[5] For related recent reviews of such proposed institutions, their 'models', and their strengths and drawbacks, see also: Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. *Bulletin of the Atomic Scientists*, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249.; Hausenloy, Jason, and Claire Dennis. 'Towards a UN Role in Governing Foundation Artificial Intelligence Models'. United Nations University - Centre for Policy Research, 19 July 2023. https://unu.edu/cpr/working-paper/towards-un-role-governing-foundation-artificial-intelligence-models. (Part II - Assessment of Proposed International Institutions in AI Governance'). Pg. 17-27 (reviewing specifically the IAEA, CERN, ICAO and IPCC models). For other, more general reviews of developments and levers in international AI governance, see also: Veale, Michael, Kira Matus, and Robert Gorwa. 'AI and Global Governance: Modalities, Rationales, Tensions'. *Annual Review of Law and Social Science* 19, no. 1 (2023). https://doi.org/10.1146/annurev-lawsocsci-020223-040749. Maas, Matthijs, Transformative AI Governance: A Literature Review. Legal Priorities Project, AI Foundations Report #3. (forthcoming 2023). For a broad research agenda into the global governance of AI, see also: Tallberg, Jonas, Eva Erman, Markus Furendal, Johannes Geith, Mark Klamberg, and Magnus Lundgren. 'The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research'. *International Studies Review* 25, no. 3 (1 September 2023): viad040. https://doi.org/10.1093/isr/viad040.

[6] For a distinct, (2x2) taxonomy of multilateral governance initiatives to AI, distinguishing between (1) initiatives that are state-led vs. non-state-led, and between (2) initiatives embedded in the existing governance architecture vs. those that establish new instruments, see also Schmitt, Lewin. 'Mapping Global AI Governance: A Nascent Regime in a Fragmented Landscape'. *AI and Ethics*, 17 August 2021. https://doi.org/10.1007/s43681-021-00083-y. For a more general taxonomy of types of legal uncertainties created by new technologies, and the resulting differences in regulatory responses, see: Crootof, Rebecca, and B. J. Ard. 'Structuring Techlaw'. *Harvard Journal of Law & Technology* 34, no. 2 (2021): 347–417. https://jolt.law.harvard.edu/assets/articlePDFs/v34/1.-Crootof-Ard-Structuring-Techlaw.pdf; Maas, Matthijs M. 'International Law Does Not Compute: Artificial Intelligence and The Development, Displacement or Destruction of the Global Legal Order'. *Melbourne Journal of International Law* 20, no. 1 (2019): 29–56.; https://law.unimelb.edu.au/__data/assets/pdf_file/0005/3144308/Maas.pdf.

(1) Rely on unilateral extraterritorial regulation. The extraterritorial approach foregoes (or at least does not prioritize) the multilateral pursuit of international regimes, norms or institutions. Rather, it aims to enact effective domestic regulations on AI developments, and then rely on the direct or extraterritorial effects of such regulations to affect the conditions or standards for AI governance in other jurisdictions. As such, this approach includes proposals to first regulate AI within (key) countries, whether by existing laws,[7] through new laws or standards developed by existing institutions, or through new domestic institutions (such as a US 'AI Control Council'[8] or a National Algorithms Safety Board[9]). These national policy levers[10] can unilaterally affect the global approach to AI, either directly–for instance, through the effect of export controls on chokepoints in the AI chip supply chains[11]–or because of the way that such regulations can spill over to other jurisdictions, as seen in discussions of a 'Brussels Effect', a 'California Effect', or even a 'Beijing Effect'.[12]

[7] Gutierrez, Carlos Ignacio. 'The Unforeseen Consequences of Artificial Intelligence (AI) on Society: A Systematic Review of Regulatory Gaps Generated by AI in the U.S.' Thesis, Pardee RAND Graduate School, 2020. https://www.rand.org/pubs/rgs_dissertations/RGSDA319-1.html.

[8] Korinek, Anton. 'Why We Need a New Agency to Regulate Advanced Artificial Intelligence: Lessons on AI Control from the Facebook Files'. *Brookings* (blog), 8 December 2021. https://www.brookings.edu/research/why-we-need-a-new-agency-to-regulate-advanced-artificial-intelligence-lessons-on-ai-control-from-the-facebook-files/.

[9] Shneiderman, Ben. 'Do We Need a National Algorithms Safety Board?' Text. *The Hill* (blog), 28 February 2023. https://thehill.com/opinion/technology/3876569-do-we-need-a-national-algorithms-safety-board/

[10] See for instance: Fischer, Sophie-Charlotte, Jade Leung, Markus Anderljung, Cullen O'Keefe, Stefan Torges, Saif M. Khan, Ben Garfinkel, and Allan Dafoe. 'AI Policy Levers: A Review of the U.S. Government's Tools to Shape AI Research, Development, and Deployment'. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, March 2021. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/AI-Policy-Levers-A-Review-of-the-U.S.-Governments-tools-to-shape-AI-research-development-and-deployment-%E2%80%93-Fischer-et-al.pdf

[11] Barbe, Andre, and Will Hunt. 'Preserving the Chokepoints: Reducing the Risks of Offshoring Among U.S. Semiconductor Manufacturing Equipment Firms'. *Center for Security and Emerging Technology*, May 2022. https://cset.georgetown.edu/publication/preserving-the-chokepoints/. Though note that in many cases, even such 'unilateral' approaches may still involve some multilateral or minilateral cooperation with selected allied states. See for instance Flynn, Carrick, and Khan. 'Multilateral Controls on Hardware Chokepoints'. *Center for Security and Emerging Technology* (blog), September 2020. https://cset.georgetown.edu/publication/multilateral-controls-on-hardware-chokepoints/.

[12] See for example Siegmann, Charlotte, and Markus Anderljung. 'The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market'. Centre for the Governance of AI, August 2022. https://www.governance.ai/research-paper/brussels-effect-ai ; Josephson, Henry. 'A California Effect for Artificial Intelligence', 2022. https://www.henryjos.com/p/a-california-effect-for-artificial.html.; Erie, Matthew S, and Thomas Streinz. 'The Beijing Effect: China's "Digital Silk Road" as Transnational Data Governance'. *New York University Journal of International Law and Politics*, 2021, 61. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3810256 ; For general account of how states can aim to pursue the global regulation of digital technologies from a domestic regulatory perspective, see also Beaumier, Guillaume, Kevin Kalomeni, Malcolm Campbell‐Verduyn, Marc Lenglet, Serena Natile, Marielle Papin, Daivi Rodima‐Taylor, Arthur Silve, and Falin Zhang. 'Global Regulations for a Digital Economy: Between New and Old Challenges'. *Global Policy* 11, no. 4 (September 2020): 515–22. https://doi.org/10.1111/1758-5899.12823.

(2) Apply existing international institutions, regimes, or norms to AI. The norm application-focused approach argues that because much of international law establishes broad, technology-neutral principles and obligations, and many domains are already subject to a wide set of overlapping institutional activities, AI technology is in fact already adequately regulated in international law.[13] As such, AI governance does not need new institutions or novel institutional models; rather the aim is to reassert, reapply, extend, and clarify long-existing international institutions and norms. This is one approach that has been taken (with greater and lesser success) to address the legal gaps initially created by some past technologies, such as submarine warfare,[14] cyberwar,[15] or data flows within the digital economy,[16] amongst others. This also corresponds to the approach taken by many international legal scholars, who argue that States should simply recognize that AI is already covered and regulated by existing norms and doctrines in international law, such as the principles of International Human Rights Law,[17] International Humanitarian Law, International Criminal Law,[18] the doctrine of state responsibility,[19] or other regimes.[20]

[13] For an argument suggesting that many types of existential risk, including transformative AI, already do receive considerable normative coverage under international law (without arguing that this should preclude the establishment of new institutions), see also the forthcoming paper: Villalobos, José Jaime, Matthijs Maas, and Christoph Winter. 'States Must Mitigate Existential Risk under International Law', forthcoming 2023.

[14] Crootof, Rebecca. 'Jurisprudential Space Junk: Treaties and New Technologies'. In Resolving Conflicts in the Law, edited by Chiara Giorgetti and Natalie Klein, 106–29, 2019. https://brill.com/view/book/edcoll/9789004316539/BP000015.xml.

[15] Eichensehr, Kristen E. 'Cyberwar & International Law Step Zero'. Texas International Law Journal 50, no. 2 (2015): 357–80. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2611198

[16] Notably, such flows were not provided for within the General Agreement on Trade in Services (GATS); however data localization measures and data flow restrictions became covered within international trade law by the WTO Appellate Body taking an evolutionary approach to interpreting GATS. See Panel Report, United States - Measures Affecting the Cross-Border Supply of Gambling and Betting Services, WT/DS285/R, adopted 10 November 2004, para. 6.287 (*US-Gambling*). See also: Mishra, Neha. 'International Trade Law Meets Data Ethics: A Brave New World'. New York University Journal of International Law and Politics 53:2 (2021), . https://doi.org/10.2139/ssrn.3689412. p. 336.

[17] Vöneky, Silja. 'How Should We Regulate AI? Current Rules and Principles as Basis for "Responsible Artificial Intelligence"', 19 May 2020. https://papers.ssrn.com/abstract=3605440; McGregor, Lorna, Daragh Murray, and Vivian Ng. 'International Human Rights Law as a Framework for Algorithmic Accountability'. International & Comparative Law Quarterly 68, no. 2 (April 2019): 309–43. https://doi.org/10.1017/S0020589319000046. See also Chinen, Mark. *The International Governance of Artificial Intelligence*. Northampton: Edward Elgar Publishing, 2023. Chapter 10.

[18] Burri, Thomas. 'International Law and Artificial Intelligence'. *German Yearbook of International Law* 60 (27 October 2017): 91–108. http://dx.doi.org/10.2139/ssrn.3060191

[19] Boutin, Bérénice. 'State Responsibility in Relation to Military Applications of Artificial Intelligence'. *Leiden Journal of International Law* 36, no. 1 (March 2023): 133–50. https://doi.org/10.1017/S0922156522000607.

[20] Though for a contrary review of existing norms, arguing that transformative AI is mostly uncovered by existing regimes in international law, see also Kemp, Luke, and Catherine Rhodes. 'The Cartography of Global Catastrophic Governance'. Global Challenges Foundation, 2020. https://globalchallenges.org/the-cartography-of-global-catastrophic-governance/.

(3) Adapt existing international institutions or norms to AI. This approach concedes that AI technology is not yet adequately or clearly governed under international law, but holds that existing international institutions could still be adapted to take on this role, and may already be doing so. This approach includes proposals that center on mapping, supporting and extending the existing AI-focused activities of existing international regimes and institutions such as the IMO, ICAO, ITU,[21] various UN agencies,[22] or other international organizations.[23] Others explore proposals for refitting existing institutions, such as expanding the G20 with a Coordinating Committee for the Governance of Artificial Intelligence',[24] or changing the mandate or composition of UNESCO's International Research Centre of Artificial Intelligence (ICRAI) or the International Electrotechnical Commission (IEC),[25] to take up a stronger role in AI governance. Finally, others explore how either States (through Explanatory Memoranda or treaty reservations), or treaty bodies (through Working Party Resolutions) could adapt existing treaty regimes to more clearly cover AI systems.[26] The emphasis here is on a 'decentralized but coordinated' approach that supports institutions to adapt to AI,[27] rather than necessarily aiming to establish new institutions in an already-crowded existing international 'regime complex'.[28]

(4) Create new international institutions to regulate AI based on the model of past or existing institutions. The institution-re-creating approach argues that AI technology does need new, distinct international institutions to be adequately governed. However, in developing designs, or making the case for such institutions, it often points

[21] See Kunz, Martina, and Seán Ó hÉigeartaigh. 'Artificial Intelligence and Robotization'. In *Oxford Handbook on the International Law of Global Security*, edited by Robin Geiss and Nils Melzer. Oxford University Press, 2021. https://papers.ssrn.com/abstract=3310421. See also

[22] Garcia, Eugenio V. 'Multilateralism and Artificial Intelligence: What Role for the United Nations?' In *The Global Politics of Artificial Intelligence*, edited by Maurizio Tinnirello, 18. Boca Raton: CRC Press, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3779866.

[23] Kunz, Martina. 'AI and International Organizations'. Accessed 31 October 2022. https://globalaigov.org/participants/igos.html.

[24] Jelinek, Thorsten, Wendell Wallach, and Danil Kerimi. 'Policy Brief: The Creation of a G20 Coordinating Committee for the Governance of Artificial Intelligence'. AI and Ethics, 6 October 2020. https://doi.org/10.1007/s43681-020-00019-y.

[25] Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. Bulletin of the Atomic Scientists, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249. Pg. 311.

[26] Smith, Bryant Walker. 'New Technologies and Old Treaties'. AJIL Unbound 114 (ed 2020): 152–57. https://doi.org/10.1017/aju.2020.28.

[27] Roberts, Huw. 'Opinion—A New International AI Body Is No Panacea'. E-International Relations (blog), 11 August 2023. https://www.e-ir.info/2023/08/11/opinion-a-new-international-ai-body-is-no-panacea/.

[28] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. Global Policy 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890. Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Should Artificial Intelligence Governance Be Centralised?: Design Lessons from History'. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 228–34. New York NY USA: ACM, 2020. https://doi.org/10.1145/3375627.3375857.

to the precedent of past or existing international institutions and regimes that have a similar model.

(5) Create entirely novel international institutional models to regulate AI. This approach argues not only that AI technology needs new international institutions, but also that past or existing international institutions (mostly) do not provide adequate models to narrowly follow or mimic.[29] This is potentially reflected in some especially ambitious proposals for comprehensive global AI regimes, or suggestions to introduce entirely new mechanisms (e.g. 'regulatory markets'[30]) to governance.

In this review we specifically focus on proposals for international AI governance and regulation that involve the creation of new international institutions for AI. That is to say, **our main focus is on approach (#4)** and, to a lesser, extent approach (#5).

We focus on new institutions as they might be better positioned to respond to the novelty, stakes and technical features of advanced AI systems.[31] Indeed, the current climate of global attention on AI seems potentially more supportive of the establishment of new landmark institutions for AI, than has been the case in past years. As AI capabilities progress at an unexpected rate, multiple government representatives and entities,[32] as well as international organizations,[33] have recently stated their support towards a new international AI governance institution. Additionally, the idea of

[29] See for instance Hausenloy and Dennis. 'Towards a UN Role in Governing Foundation Artificial Intelligence Models'. Pg. 3 ("International AI governance cannot be achieved by copy-pasting existing models, but rather by using these historical examples to employ a multi-pronged approach").

[30] Hadfield, Gillian K, and Jack Clark. 'Regulatory Markets: The Future of AI Governance', April 2023. https://arxiv.org/ftp/arxiv/papers/2304/2304.04914.pdf.

[31] See Trager, Robert, and others. 'International Governance of Civilian AI: A Jurisdictional Certification Approach'. arXiv, 29 August 2023. https://doi.org/10.48550/arXiv.2308.15514. pg. 11-14.

[32] See for example: National Security Commission on Artificial Intelligence. 'Final Report'. National Security Commission on Artificial Intelligence, March 2021. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. (Chapter 15). (United States); Stacey, Kiran. 'UK Should Play Leading Role on Global AI Guidelines, Sunak to Tell Biden'. *The Guardian*, 31 May 2023, sec. Technology. https://www.theguardian.com/technology/2023/may/31/uk-should-play-leading-role-in-developing-ai-global-guidelines-sunak-to-tell-biden; UN Press. 'International Community Must Urgently Confront New Reality of Generative, Artificial Intelligence, Speakers Stress as Security Council Debates Risks, Rewards'. UN Press, 18 July 2023. https://press.un.org/en/2023/sc15359.doc.htm.

[33] Committee on Artificial Intelligence (CAI). 'Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law'. Council of Europe, 6 January 2023. https://rm.coe.int/cai-2023-01-revised-zero-draft-framework-convention-public/1680aa193f.; Guterres, António. 'Secretary-General's Remarks to the Security Council on Artificial Intelligence'. United Nations Secretary-General, 18 July 2023. https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence.

establishing such institutions has taken root among many of the leading actors in the AI industry.[34]

With this, our review comes with two caveats. In the first place, our focus on this institutional approach above others does not mean that pursuing the creation of new institutions is necessarily an easy strategy, or more feasible than the other approaches listed above. Indeed, proposals for new treaty regimes or international institutions for AI–especially when they draw analogies with organizations that were set up decades ago–may often underestimate how much the ground of global governance has changed in recent years. As such, they do not always reckon fully with the strong trends and forces in global governance which, for better or worse, have come to frequently push States towards relying on the extension of existing norms (approach #2) or the adaptation of existing institutions (approach #3),[35] rather than creating novel institutions. Likewise, there are further trends that push towards a shift in US policy, towards pursuing international cooperation through nonbinding international agreements rather than treaties;[36] as well as concerns that by some trends, international organizations may be taking up a less central role in international relations today than they have in the past.[37] All of these trends should temper, or at least inform, proposals to establish new institutions.

Furthermore, even if one is determined to pursue the establishment of a new international institution along one of the models discussed here, many key open

[34] Altman, Sam, Greg Brockman, and Ilya Sutskever. 'Governance of Superintelligence'. OpenAI, 22 May 2023. https://openai.com/blog/governance-of-superintelligence; Ho, Lewis, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, et al. 'International Institutions for Advanced AI'. arXiv, 10 July 2023. https://doi.org/10.48550/arXiv.2307.04699.

[35] See for instance: Eichensehr, Kristen E. 'Cyberwar & International Law Step Zero'. *Texas International Law Journal* 50, no. 2 (2015): 357–80.; and also Alter, Karen J., and Kal Raustiala. 'The Rise of International Regime Complexity'. *Annual Review of Law and Social Science* 14, no. 1 (2018): 329–49. https://doi.org/10.1146/annurev-lawsocsci-101317-030830. Pg 337 ("[g]lobal governance solutions [...] must take one of two approaches: (a) International actors can attempt to create an encompassing regime that can address all dimensions of the problem, or (b) international actors can accept that policy solutions will be crafted, coordinated, and implemented within a larger regime complex. [...] although the first option might be more efficient and effective, it is rarely the solution adopted").

For a discussion of seven global trends that have driven regime complexity and fragmentation in global governance broadly (institutional density, accretion, state power shifts, state preference changes, modernity, demands for representation and voice, and preference for local governance responses), and how these might apply in the context of global AI governance, see: Maas, Matthijs M. 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. University of Copenhagen, 2020. http://www.legalpriorities.org/documents/Maas-PhD-Dissertation.pdf. Pg. 286-291.

[36] Bradley, Curtis, Jack Landman Goldsmith, and Oona A. Hathaway. 'The Rise of Nonbinding International Agreements: An Empirical, Comparative, and Normative Analysis'. *The University of Chicago Law Review* 90, no. 5 (2023). https://doi.org/10.2139/ssrn.4023641.

[37] Debre, Maria J., and Hylke Dijkstra. 'Are International Organisations in Decline? An Absolute and Relative Perspective on Institutional Change'. Global Policy 14, no. 1 (2023): 16–30. https://doi.org/10.1111/1758-5899.13170.

questions remain about the optimal route to design and establish that organization, including: (1) Given that many institutional functions might be required to adequately govern advanced AI systems, might there be a need for 'hybrid' or combined institutions with a dual mandate, like the IAEA?[38] (2) Should an institution be tightly centralized or could it be a relatively decentralized, with one or more new institutions orchestrating the AI policy activities of a constellation of many other (existing or new) organizations?[39] (3) Should such an organization be established formally, or are informal club approaches adequate in the first instance?[40] (4) Should voting rules within such institutions work on the grounds of consensus, or simple majority? (5) What rules should govern the adaptation or updating of the institution's mission and mandate, to track ongoing developments in AI? This review will briefly flag and discuss some of these questions in Part II, but will leave many of them open for future research.

Regarding terminology, we will use both 'international institution' and 'international organization' interchangeably, and broadly to refer to any of (1) formally established formal Intergovernmental Organizations (FIGOs) founded through a constituent document (e.g. WTO, WHO); (2) treaty bodies or secretariats that have a more limited mandate, primarily supporting the implementation of a treaty or regime (e.g. BWC Implementation Support Unit); and (3) 'informal IGOs' (IIGOs) that consist of loose 'task

[38] We thank Harry Law for this observation. See also the discussion of research Direction 4, in Part II.

[39] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890.

[40] Morin, Jean-Frédéric, Hugo Dobson, Claire Peacock, Miriam Prys-Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, et al. 'How Informality Can Address Emerging Issues: Making the Most of the G7'. *Global Policy* 10, no. 2 (May 2019): 267–73. https://doi.org/10.1111/1758-5899.12668.

groups' and coalitions of states (e.g. the G7, BRICS, G20).[41] We use 'model' to refer to the general cluster of institutions under discussion; we will use 'function' to refer to a given institutional model's purpose or role. We use 'AI proposals' to refer to the precise institutional models that are proposed for international AI governance.

I. Review of Institutional Models

Below, we review a range of institutional models that have been proposed for AI governance. For each model, we discuss their general functions, different variations or forms, a range of examples that are frequently invoked, and explicit AI governance proposals that follow this model. In addition, we will highlight additional examples that have not received much attention but that we believe could be promising. Finally, where applicable, we will highlight existing critiques of a given model.

Model 1: Scientific consensus-building

1.1 Functions and types: The functions of the scientific consensus-building institutional model are to (1) increase general policymaker and public awareness of an issue; and especially to (2) establish a scientific consensus on an issue. The aim of this is to facilitate greater common knowledge or shared perception of an issue amongst States, with the aim to motivate national action, or enable international agreements. Overall, their goal is not to establish an international consensus on how to respond, or to hand down regulatory recommendations directly, but simply to provide a basic knowledge base to underpin the decisions of key actors. By design they are, or aim, to be

[41] For definitions and distinctions of IGO, FIGO, and IIGOs in the context of proposals for an international AI governance institution, see also: Erdélyi, Olivia J., and Judy Goldsmith. 'Regulating Artificial Intelligence: Proposal for a Global Solution'. Government Information Quarterly 39, no. 4 (1 October 2022): 101748. https://doi.org/10.1016/j.giq.2022.101748. pg. 12.

> ("we define an IGO as a formal entity (1) established by an international agreement governed by international law; (2) with at least three (sometimes two) members — typically states but increasingly also IGOs; and (3) having at least one organ with a will distinct from that of its members. FIGOs' organizational purpose is laid down in a binding international agreement such as a treaty or a formal legal act of another IGO, their membership is clearly defined in the founding legal act, and they have a permanent and significant institutionalization in place. By contrast, IIGOs operate based on an explicitly shared, but informal expectation about purpose, their membership is not always clear, as members are explicitly associated but only by non-legal mutual acknowledgment, and they do not possess any significant institutionalization. NGOs differ from IGOs in that they are not created by treaty — meaning they are governed by national rather than international law — and their membership is made up of non-state actors.").

For a general discussion of the growing role of IIGOs in global governance, see also: Vabulas, Felicity, and Duncan Snidal. 'Informal IGOs as Mediators of Power Shifts'. Global Policy 11, no. S3 (2020): 40–50. https://doi.org/10.1111/1758-5899.12869. On the importance of also considering IIGOs and not just FIGO's in evaluations of the contemporary role of international organizations, see also Roger, Charles B., and Sam S. Rowan. 'Analyzing International Organizations: How the Concepts We Use Affect the Answers We Get'. *The Review of International Organizations* 17, no. 3 (1 July 2022): 597–625. https://doi.org/10.1007/s11558-021-09432-2.

non-political—as in the IPCC's mantra to be "policy-relevant and yet policy-neutral, never policy-prescriptive".[42]

1.2 Common examples: Commonly cited examples of scientific consensus-building institutions include most notably the Intergovernmental Panel on Climate Change (IPCC),[43] the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES),[44] and the Scientific Assessment Panel (SAP) of the United Nations Environment Programme (UNEP).[45]

1.3 Underexplored examples: An example that has not yet been invoked but that could be promising to explore, is the Antarctic Treaty's Committee for Environmental Protection (CEP), which provides expert advice to the Antarctic Treaty Consultative Meetings and which combines scientific consensus-building models with risk-management functions, supporting the Protocol on Environmental Protection to the Antarctic Treaty.[46] Another example could be the World Meteorological Organization (WMO) which monitors weather and climatic trends and makes information available.

1.4 Proposed AI institutions along this model: There have been a range of proposals for scientific consensus-building institutions for AI. Indeed, in 2018, the precursor initiative to what would become the Global Partnership on AI (GPAI) was initially envisaged by France and Canada as an **Intergovernmental Panel on AI (IPAI)** along the IPCC model.[47] This proposal was supported by many researchers:

[42] 'IPCC - Intergovernmental Panel on Climate Change'. Accessed 28 August 2023. https://archive.ipcc.ch/organization/organization.shtml; see also Havstad, Joyce C., and Matthew J. Brown. 'Neutrality, Relevance, Prescription, and the IPCC'. Public Affairs Quarterly 31, no. 4 (2017): 303–24. https://www.jstor.org/stable/44732800.

[43] Ho, Lewis, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, et al. 'International Institutions for Advanced AI'. arXiv, 10 July 2023. https://doi.org/10.48550/arXiv.2307.04699. See also: Neufville, Robert de, and Seth D. Baum. 'Collective Action on Artificial Intelligence: A Primer and Review'. *Technology in Society* 66 (1 August 2021): 101649. https://www.sciencedirect.com/science/article/pii/S0160791X2100124X; Whitfield, Robert. 'Effective, Timely and Global: The Urgent Need for Good Global Governance of AI'. World Federalist Movement and Institute for Global Policy, 2020. https://www.wfm-igp.org/publication/effective-timely-and-global-the-urgent-need-for-good-global-governance-of-ai/., pg. 63. Mulgan, Geoff, Thomas Malone, Divya Siddarth, Saffron Huang, Joshua Tan, and Lewis Hammond. 'The Case for a Global AI Observatory (GAIO)'. Carnegie Council for Ethics in International Affairs, 6 June 2023. https://www.carnegiecouncil.org/media/article/the-case-for-a-global-ai-observatory-gaio-2023.

[44] Bak-Coleman, Joseph, Carl T. Bergstrom, Jennifer Jacquet, James Mickens, Zeynep Tufekci, and Timmons Roberts. 'Create an IPCC-like Body to Harness Benefits and Combat Harms of Digital Tech'. *Nature* 617, no. 7961 (May 2023): 462–64. https://doi.org/10.1038/d41586-023-01606-9. (referring to IPCC and IPBES). See also Mulgan, Geoff, and Divya Siddarth. 'The World Needs A Global AI Observatory'. *Noema*, 29 June 2023. https://www.noemamag.com/the-world-needs-a-global-ai-observatory.

[45] Ho and others, 'International Institutions for Advanced AI'.

[46] 'The Committee for Environmental Protection | Antarctic Treaty'. Accessed 28 August 2023. https://www.ats.aq/e/committee.html.

[47] Simonite, Tom. 'Canada, France Plan Global Panel to Study the Effects of AI'. *Wired*, 6 December 2018. https://www.wired.com/story/canada-france-plan-global-panel-study-ai/.

Kemp and others proposed an IPAI that could measure, track, and forecast progress in AI, as well as its use and impacts, in order to "provide a legitimate, authoritative voice on the state and trends of AI technologies."[48] They argue that an IPAI could perform structural assessments every three years as well as take up quick response special issue assessments. Around that time, Mialhe also proposed an IPAI model, as an institution that would gather a large and global group of experts "to inform dialogue, coordination, and pave the way for efficient global governance of AI."[49]

More recently, Ho and others have proposed an intergovernmental **Commission on Frontier AI**, to "establish a scientific position on opportunities and risks from advanced AI and how they may be managed," to help increase public awareness and understanding, to "contribute to a scientifically informed account of AI use and risk mitigation [and to] be a source of expertise for policymakers."[50] Bremmer and Suleyman have proposed a **global scientific body** to objectively advise governments and international bodies on questions as basic as what AI is and what kinds of policy challenges it poses.[51] They draw a direct link to the IPCC model, noting that "this body would have a global imprimatur and scientific (and geopolitical) independence [...] [a]nd its reports could inform multilateral and multistakeholder negotiations on AI."[52] Bak-Coleman and others have argued in favor of an **Intergovernmental Panel on Information Technology**, an independent, IPCC-like panel charged with studying the "impact of emerging information technologies on the world's social, economic, political and natural systems."[53] In their view, this panel would focus on many 'computational systems', including "search engines, online banking, social-media platforms and large language models" and would have leverage to persuade companies to share key data.[54]

Finally, Mulgan and others have recently proposed a **Global AI Observatory (GAIO)** as an institution that "would provide the necessary facts and analysis to support decision-making [and] would synthesize the science and evidence needed to support a

[48] Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[49] Miailhe, Nicolas. 'AI & Global Governance: Why We Need an Intergovernmental Panel for Artificial Intelligence'. United Nations University Centre for Policy Research, 20 December 2018. https://cpr.unu.edu/ai-global-governance-why-we-need-an-intergovernmental-panel-for-artificial-intelligence.html.

[50] Ho and others, 'International Institutions for Advanced AI', p. 2.

[51] Bremmer, Ian, and Mustafa Suleyman. 'The AI Power Paradox'. *Foreign Affairs*, 16 August 2023. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox.

[52] Ibid.

[53] Bak-Coleman and others, 'Create an IPCC-like Body to Harness Benefits and Combat Harms of Digital Tech'.

[54] Ibid.

diversity of governance responses."[55] Again drawing a direct comparison to the IPCC, they anticipate that such a body could set the foundation for more serious regulation of AI, through six activities: (1) a global standardized incident reporting database, (2) a registry of crucial AI systems, (3) a shared body of data and analysis of the key facts of the AI ecosystem; (4) working groups exploring global knowledge about the impacts of AI on critical areas; (5) the ability to offer legislative assistance and model laws; and (6) the ability to orchestrate global debate through an annual report on the state of AI.[56] They have since incorporated this proposal within a larger '**Framework for the International Governance of AI**' by the Carnegie Council for Ethics in International Affairs's Artificial Intelligence & Equality Initiative, alongside other components such as a **neutral technical organization** to analyze "which legal frameworks, best practice, and standards have risen to the highest level of global acceptance'.[57]

1.5 Critiques of this model: One concern that has been expressed is that AI governance is currently too institutionally immature to support an IPCC-like model, since, as argued by Roberts; "the IPCC [...] was preceded by almost two decades of multilateral scientific assessments, before being formalised".[58] He considers that this may be a particular problem for replicating that model for AI, given that some AI risks are currently still subject to significantly less scientific consensus.[59] Separately, Bak-Coleman and others argue that a scientific consensus-building organization for digital technologies would face a far more difficult research environment than the IPCC and IPBES because, as opposed to the rich data and scientifically well understood mechanisms that characterize climate change and ecosystem degradation, research into the impacts of digital technologies often faces data access restrictions.[60] Ho and others have argued that a Commission on Frontier AI would face more general scientific challenges in adequately studying future risks 'on the horizon', as well as potential politicization, both of which might inhibit the ability of such a body to effectively build

[55] Mulgan, Geoff, Thomas Malone, Divya Siddarth, Saffron Huang, Joshua Tan, and Lewis Hammond. 'The Case for a Global AI Observatory (GAIO)'. Carnegie Council for Ethics in International Affairs, 6 June 2023. https://www.carnegiecouncil.org/media/article/the-case-for-a-global-ai-observatory-gaio-2023. See also Mulgan, Geoff, and Divya Siddarth. 'The World Needs A Global AI Observatory'. *Noema*, 29 June 2023. https://www.noemamag.com/the-world-needs-a-global-ai-observatory.

[56] Ibid.

[57] Artificial Intelligence & Equality Initiative. 'A Framework for the International Governance of AI'. Carnegie Council for Ethics in International Affairs, 5 July 2023. https://www.carnegiecouncil.org/media/article/a-framework-for-the-international-governance-of-ai.

[58] Roberts, Huw. 'Opinion—A New International AI Body Is No Panacea'. E-International Relations (blog), 11 August 2023. https://www.e-ir.info/2023/08/11/opinion-a-new-international-ai-body-is-no-panacea/.; and see generally Nature. 'Will the World Ever See Another IPCC-Style Body?' Nature 615, no. 7950 (1 March 2023): 7–8. https://doi.org/10.1038/d41586-023-00572-6.

[59] Ibid.

[60] Bak-Coleman and others, "Create an IPCC-like Body to Harness Benefits and Combat Harms of Digital Tech'. Pg. 464.

consensus.[61] Indeed, it is possible that in the absence of decisive and incontrovertible evidence about the trajectory and risks of AI, a scientific consensus-building institution would likely struggle to deliver on its core mission, and might instead spark significant scientific contestation and disagreement amongst AI researchers instead.

Model 2: Political consensus-building and norm-setting

2.1 Functions and types: The function of political consensus-building and norm-setting institutions is to help States come to greater political agreement and convergence about the way to respond to an (usually) clearly identified and (ideally) agreed issue or phenomenon. Their aim is to reach the required political consensus necessary to either align national policymaking responses sufficiently well, achieving some level of harmonization that reduces trade restrictions or impedes progress towards addressing the issue; or to help begin negotiations on other institutions that establish more stringent regimes. Political consensus-building institutions do this by providing fora for discussion and debate that can aid the articulation of potential compromises between State interests, and by exerting normative pressure on States towards certain goals. In a norm-setting capacity, institutions can also draw on (growing) political consensus to set and share informal norms, even if formal institutions have not yet been created. For instance, if negotiations for a regulatory or control institution are held up, slowed, or fail, political consensus-building institutions can also play a norm-setting function by establishing informal standards for behavior, as soft law. While such norms are not as strictly specified, or as enforceable, as hard law regulations, they can still carry force and see takeup.

2.2 Common examples: There are a range of examples of political consensus-building institutions. Some of these are broad, such as conferences of parties to a treaty (also known as COPs, the most popular one being the United Nations Framework Convention on Climate Change's (UNFCCC)'s (COP).[62] Many others, however, such as the Organization for Economic Co-operation and Development (OECD), the G20, and the G7, reflect smaller, and at times more informal governance 'clubs', which can often move ahead towards policy-setting more quickly because their membership is already somewhat aligned,[63] and because many of them have already begun to undertake activities or incorporate institutional units focused on AI developments.[64]

[61] Ho and others, 'International Institutions for Advanced AI', p. 8-9.

[62] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890. Pg. 551.

[63] Morin, Jean‐Frédéric, Hugo Dobson, Claire Peacock, Miriam Prys‐Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, et al. 'How Informality Can Address Emerging Issues: Making the Most of the G7'. *Global Policy* 10, no. 2 (May 2019): 267–73. https://doi.org/10.1111/1758-5899.12668. For a general discussion of the 'breadth vs. depth dilemma', see also Cihon and others, 'Fragmentation and the Future: Investigating Architectures for International AI Governance', 2020, pg. 549-550.

[64] Take for instance the OECD's AI Policy Observatory: OECD. 'The OECD Artificial Intelligence Policy Observatory'. Accessed 17 September 2020. https://www.oecd.ai/.

Gutierrez and others have reviewed a range of historical cases of (domestic and global) soft law governance that they argue could provide lessons for AI. These include a range of institutional activities, such as UNESCO's 1997 Universal Declaration on the Human Genome and Human Rights, 2003 International Declaration on Human Genetic Data, and 2005 Universal Declaration on Bioethics and Human Rights,[65] the Environmental Management System (ISO 14001), the Sustainable Forestry Practices by the Sustainable Forestry Initiative and Forest Stewardship Council, and the Leadership in Energy and Environmental Design initiative.[66] Others, however, such as the Internet Corporation for Assigned Names and Numbers (ICANN), the Asilomar rDNA Guidelines, the International Gene Synthesis Consortium, the International Society for Stem Cell Research Guidelines, the BASF Code of Conduct, the Environmental Defense Fund, and the DuPont Risk Frameworks, offer greater examples of success.[67] Turner has likewise argued that the ICANN offers a model for international AI governance that manages to develop productive internet policy.[68] Elsewhere, Harding has argued that the 1967 Outer Space Treaty offered a telling case of a treaty regime that quickly crystallized State expectations and policies around safe innovation in a then-novel area of science.[69] Finally, Feijóo and others suggest that 'new technology diplomacy' on AI could involve a series of meetings or global conferences on AI, which could draw lessons from experiences such as the World Summits on the Information Society (WSIS).[70]

2.3 Underexplored examples: Examples of norm-setting institutions that formulate and share relevant soft-law guidelines on technology include the International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), the International Telecommunication Union (ITU), or the United Nations Commission on International Trade Law (UNCITRAL)'s Working Group on Electronic Commerce.[71] Another good example of a political consensus-building and norm-setting

[65] See also Stevens, Yvonne A. 'Soft Law Governance: A Historical Perspective from Life-Science Technologies', 61 JURIMETRICS J. (2020). https://lsi.asulaw.org/softlaw/wp-content/uploads/sites/7/2021/04/121-131-stevens-special-issue-article.pdf

[66] Gutierrez, Carlos Ignacio, Gary E. Marchant, and Lucille Tournas. 'Lessons for Artificial Intelligence from Historical Uses of Soft Law Governance'. *JURIMETRICS* 61, no. 1 (29 December 2020). https://doi.org/10.2139/ssrn.3775271.

[67] Ibid.

[68] Turner, Jacob. *Robot Rules: Regulating Artificial Intelligence*. New York, NY: Springer Berlin Heidelberg, 2018.pg 240-242.

[69] See also Harding, Verity. 'Lessons from History: What Can Past Technological Breakthroughs Teach the AI Community Today', 2020. https://www.bennettinstitute.cam.ac.uk/blog/lessons-history-what-can-past-technological-breakt/

[70] Feijóo, Claudio, Youngsun Kwon, Johannes M. Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. 'Harnessing Artificial Intelligence (AI) to Increase Wellbeing for All: The Case for a New Technology Diplomacy'. Telecommunications Policy 44, no. 6 (6 May 2020): 101988. https://doi.org/10.1016/j.telpol.2020.101988. Pg. 12.

[71] United Nations Commission On International Trade Law. 'Working Group IV: Electronic Commerce'. Accessed 18 September 2023.

initiative could be found in the 1998 Lysøen Declaration,[72] an initiative by Canada and Norway that expanded to 11 highly committed States along with several NGOs, and which kicked off a 'Human Security Network' that achieved a significant and outsized global impact, including the Ottawa Treaty ban on antipersonnel mines; the Rome Treaty establishing the International Criminal Court; the Kimberley process aimed at inhibiting the flow of conflict diamonds; and landmark Security Council resolutions on Children and Armed Conflict and Women, Peace and Security. Another norm-setting institution that is not yet often invoked in AI discussions but that could be promising to explore, is the Codex Alimentarius Commission (CAC), which develops and maintains the Food and Agriculture Organization (FAO)'s Codex Alimentarius, a collection of non-enforceable but internationally recognized standards and codes of practice about various aspects of food production, food labeling, and safety. Another example of a 'club' under this model which is not often mentioned but that could be influential is the BRICS group, which recently expanded from five to eleven members.

2.4 Proposed AI institutions along this model: Many proposals for political consensus-building institutions on AI tend to not focus on the establishment of new institutions, arguing instead that it is best to put AI issues on the agenda of existing (established and recognized) consensus-building institutions (e.g. the G20), or of existing norm-setting institutions (e.g. ISO). Indeed, even recent proposals for new international institutions still emphasize that these should link up well with already-ongoing initiatives, such as the G7 Hiroshima Process on AI.[73]

However, there have been proposals for new political consensus-building institutions. Erdelyi and Goldsmith proposed an **International AI Organisation (IAIO)**, "to serve as an international forum for discussion and engage in standard setting activities".[74] They argue that "at least initially, the IAIO should start out as an IIGO displaying a relatively low level of institutional formality and using soft law instruments, such as non-binding recommendations, guidelines, and standards, to support national policymakers in the conception and design of AI-related regulatory policies."[75] Moreover, they emphasize that the IAIO 'should be hosted by a neutral country to provide for a safe environment, limit avenues for political conflict, and build a climate of mutual

https://uncitral.un.org/en/working_groups/4/electronic_commerce. We thank Matteo Pistillo for this suggestion.

[72] Maas, 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. pg. 308. See also Basu, Arindrajit, and Justin Sherman. 'Two New Democratic Coalitions on 5G and AI Technologies'. Lawfare, 6 August 2020. https://www.lawfareblog.com/two-new-democratic-coalitions-5g-and-ai-technologies.

[73] As directly referred to by Ho and others, 'International Institutions for Advanced AI', Pg. 2.

[74] Erdélyi, Olivia J., and Judy Goldsmith. 'Regulating Artificial Intelligence: Proposal for a Global Solution'. *Government Information Quarterly* 39, no. 4 (1 October 2022): 101748. https://doi.org/10.1016/j.giq.2022.101748. And see previously Erdelyi, Olivia J, and Judy Goldsmith. 'Regulating Artificial Intelligence: Proposal for a Global Solution'. In *Proceedings of the 2018 AAAI / ACM Conference on Artificial Intelligence, Ethics and Society*, 95–101, 2018. https://dl.acm.org/doi/10.1145/3278721.3278731.

[75] Ibid. pg. 14.

tolerance and appreciation.'[76] More recently, the US's National Security Commission on Artificial Intelligence's final report included a proposal for an **Emerging Technology Coalition**, "to promote the design, development, and use of emerging technologies according to democratic norms and values; coordinate policies and investments to counter the malign use of these technologies by authoritarian regimes; and provide concrete, competitive alternatives to counter the adoption of digital infrastructure made in China."[77] Recently, Marcus and Reuel have also proposed the creation of an '**International Agency for AI (IAAI)**', tasked with convening experts and developing tools to help find "governance and technical solutions to promote safe, secure and peaceful AI technologies."[78]

At the looser organizational end, Feijóo and others have proposed a **new technology diplomacy initiative** as 'a renewed kind of international engagement aimed at transcending narrow national interests and seeks to shape a global set of principles.' In their view such a framework could 'lead to an international constitutional charter for AI.'[79] Finally, Jernite and others have proposed a multi-party international **Data Governance Structure**, a multi-party, distributed governance arrangement for improving the global systematic and transparent management of language data at a global level, and which includes a **Data Stewardship Organization** in order to develop 'appropriate management plans, access restrictions, and legal scholarship'.[80] Other proposed organizations are also more focused on supporting states in implementing AI policy, such as through training. For instance, Turner has proposed creating an **International Academy for AI Law and Regulation**.[81]

2.5 Critiques of this model: There have not generally been many in-depth critiques of proposals for new political consensus-building or norm-setting institutions. However, some concerns that have been raised focus on the difficult tradeoffs that consensus-building institutions face in deciding whether to prioritize breadth of

[76] Ibid.

[77] National Security Commission on Artificial Intelligence. 'Final Report'. National Security Commission on Artificial Intelligence, March 2021. https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. (Chapter 15).

[78] The Economist. 'The World Needs an International Agency for Artificial Intelligence, Say Two AI Experts'. *The Economist*, 18 April 2023. https://www.economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts.

[79] Feijóo, Claudio, Youngsun Kwon, Johannes M. Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. 'Harnessing Artificial Intelligence (AI) to Increase Wellbeing for All: The Case for a New Technology Diplomacy'. *Telecommunications Policy* 44, no. 6 (6 May 2020): 101988. https://doi.org/10.1016/j.telpol.2020.101988. Pg. 11.

[80] Jernite, Yacine, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, et al. 'Data Governance in the Age of Large-Scale Data-Driven Language Technology'. In 2022 ACM Conference on Fairness, Accountability, and Transparency, 2206–22, 2022. https://doi.org/10.1145/3531146.3534637.

[81] Turner, Jacob. Robot Rules: Regulating Artificial Intelligence. New York, NY: Springer Berlin Heidelberg, 2018. Pg. 254.

membership and inclusion or depth of mission alignment. Institutions that aim to foster consensus across a very broad swath of actors may be very slow to reach such normative consensus, and even when they do may only achieve a 'lowest-common-denominator' agreement.[82] On the other hand, others have countered that AI consensus-building institutions or fora will need to be sufficiently inclusive—in particular, and possibly controversially, with regards to China[83]—if they do not want to run the risk of producing a fractured and ineffective regime, or even see negotiations implode over the political question of who was invited or excluded.[84] Finally, a more foundational challenge to political consensus-building institutions is that while it may result in (the appearance of) joint narratives, this may not have much teeth if the agreement is not binding.[85]

Model 3: Coordination of policy and regulation

3.1 Functions and types: The functions of this institutional model are to help align and coordinate policies, standards, or norms,[86] in order to ensure a coherent international approach to a common problem. There is significant internal variation in the set-up of institutions under this model, with various subsidiary functions. For instance, such institutions may: (1) *directly regulate* the deployment of a technology in relative detail, requiring States to comply and implement those regulations at the national level; (2) assist States in the national *implementation* of agreed AI policies; (3) focus on the *harmonization and coordination* of policies; (4) focus on the *certification* of industries or jurisdictions to ensure they comply with certain standards; or (5) in some cases, take up functions related to *monitoring and enforcing* norm compliance.

3.2 Common examples: Common examples of policy-setting institutions include the World Trade Organization (WTO) as an example of an empowered, centralized regulatory institution.[87] Other examples given of regulatory institutions include the

[82] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. Global Policy 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890. Pg 550.

[83] Roberts, Huw. 'Letter: Why Excluding China from the AI Summit Would Be a Mistake'. Financial Times, 21 August 2023. https://www.ft.com/content/3829707c-b93e-4715-bc7e-4de917e76914. But for a critical response see Chalmers, Alex, and Nathan Benaich. 'China Has No Place at the AI Safety Summit', 31 August 2023. https://www.airstreet.com/blog/china-ai-safety-summit.

[84] For instance, it has been argued that this was one factor that may have derailed the progress of the Nuclear Security Summits. Stover, Dawn. 'The Controversial Legacy of the Nuclear Security Summit'. Bulletin of the Atomic Scientists (blog), 4 October 2018. https://thebulletin.org/2018/10/the-controversial-legacy-of-the-nuclear-security-summit/.

[85] We thank Harry Law for this point.

[86] See also Stix, Charlotte. 'Foundations for the Future: Institution Building for the Purpose of Artificial Intelligence Governance'. *AI and Ethics* 2, no. 3 (1 August 2022): 463–76. https://doi.org/10.1007/s43681-021-00093-w.

[87] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890. Pg. 547. Jordan, Richard, Nicholas Emery-Xu, and Robert Trager. 'International Governance of Artificial Intelligence', (working paper); Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of

International Civil Aviation Organization (ICAO), the International Maritime Organization (IMO), the International Atomic Energy Agency (IAEA), and the Financial Action Task Force (FATF).[88] Examples of policy-coordinating institutions may include the United Nations Environment Programme (UNEP), which synchronized international agreements on the environment, and facilitated new agreements including the 1985 Vienna Convention for the Protection of the Ozone Layer.[89] Nemitz has pointed to the example of the institutions created under the United Nations Convention on the Law of the Sea (UNCLOS) as a model for an AI regime, including an international court to enforce the proposed treaty."[90] Finally, Sepasspour has proposed the establishment of an 'AI Ethics and Safety Unit' within the International Electrotechnical Commission (IEC), a model that is 'inspired by the Food and Agriculture Organization's (FAO) Food Safety and Quality Unit and Emergency Prevention System for Food Safety early warning system'.[91]

3.3 Underexplored examples: Examples that are not yet often discussed, but that could be useful or insightful, include the European Monitoring and Evaluation Programme (EMEP), which implements the 1983 Convention on Long-Range Transboundary Air Pollution—a regime that has proven particularly adaptive.[92] A more sui generis example is that of international financial institutions, like the World Bank or the International Monetary Fund (IMF), which tend to shape domestic policy indirectly through conditional access to loans or development funds.

Artificial Intelligence'. *Bulletin of the Atomic Scientists*, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249. Pg. 305.

[88] Ho and others, 'International Institutions for Advanced AI', p. 2; see also the reference to the IAEA in Chowdhury, Rumman. 'AI Desperately Needs Global Oversight'. *Wired*, 6 April 2023. https://www.wired.com/story/ai-desperately-needs-global-oversight/; as well as: Trager, Robert and others. 'International Governance of Civilian AI: A Jurisdictional Certification Approach'. arXiv, 29 August 2023. https://doi.org/10.48550/arXiv.2308.15514. (referring to the models of the ICAO, IMO, and FATF); Feijóo, Claudio, Youngsun Kwon, Johannes M. Bauer, Erik Bohlin, Bronwyn Howell, Rekha Jain, Petrus Potgieter, Khuong Vu, Jason Whalley, and Jun Xia. 'Harnessing Artificial Intelligence (AI) to Increase Wellbeing for All: The Case for a New Technology Diplomacy'. *Telecommunications Policy* 44, no. 6 (6 May 2020): 101988. https://doi.org/10.1016/j.telpol.2020.101988. Pg. 12. (referring to FATF).

[89] Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[90] Nemitz, Paul. 'Fundamentals of International Law on AI'. In Remaking the World: Toward an Era of Global Enlightenment, edited by Nguyen Anh Tuan. Boston Global Forum/United Nations Academic Impact, 2021. https://bostonglobalforum.org/publications/the-age-of-global-enlightenment/.

[91] Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. Bulletin of the Atomic Scientists, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249. pg. 311.

[92] Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (November 2020): 545–56. https://doi.org/10.1111/1758-5899.12890. pg . 550.

3.4 Proposed AI institutions along this model: Specific to advanced AI, recent proposals for regulatory institutions include Ho and other's **Advanced AI Governance Organisation**, which "could help internationalize and align efforts to address global risks from advanced AI systems by setting governance norms and standards, and assisting in their implementation."[93]

Trager and others have proposed an **International AI Organization (IAIO)** to certify jurisdictions' compliance with international oversight standards. These would be enforced through a system of conditional market access in which trade barriers would be imposed on jurisdictions which are not certified or whose supply chains integrate AI from non-IAIO certified jurisdictions. Among other advantages, the authors suggest this system could be less vulnerable to proliferation of industry secrets by having States establish their own domestic regulatory entities, rather than having international jurisdictional monitoring (as is the case with the IAEA). However, the authors also propose that the IAIO could provide monitoring services to governments that have not yet built their own monitoring capabilities. The authors argue their model has several advantages over others, including agile standard-setting, monitoring and enforcement.[94]

In a regional context, Stix has proposed an **EU AI Agency** which, among other roles, could be an analyser of gaps in AI policy and a developer of policies that could fill that gap. For this agency to be effective, Stix suggests it should be independent from political agendas by, for instance, having a mandate that does not coincide with election cycles.[95] Webb has proposed a '**Global Alliance on Intelligence Augmentation' (GAIA)**, which would bring together experts from different fields to set best practices for AI.[96]

Chowdhury has proposed a **generative AI global governance body**, as a "consolidated ongoing effort with expert advisory and collaborations [which] should receive advisory input and guidance from industry, but have the capacity to make independent binding decisions that companies must comply with.'[97] In her analysis, this body should be funded via unrestricted and unconditional funds by all AI companies engaged in the creation or use of generative AI and it should "cover all aspects of generative AI models, including their development, deployment, and use as it relates to the public good. It should build upon tangible recommendations from civil society and academic organizations, and have the authority to enforce its decisions, including the power to require changes in the design or use of generative AI models, or even halt their use altogether if necessary."[98]

[93] Ho and others, 'International Institutions for Advanced AI'.

[94] Trager and others, "International Governance of Civilian AI'.

[95] Stix, "Foundations for the Future: Institution Building for the Purpose of Artificial Intelligence Governance'.

[96] Amy Webb, 'The Big Nine: How the Tech Titans and their Thinking Machines Could Warp Humanity' (Public Affairs, 2019); https://www.politico.eu/article/build-democracy-into-ai-combat-china/.

[97] Chowdhury, Rumman. 'AI Desperately Needs Global Oversight'. *Wired*, 6 April 2023. https://www.wired.com/story/ai-desperately-needs-global-oversight/.

[98] Ibid.

A proposal for a policy-coordinating institution is Kemp and others' **Coordinator and Catalyser of International AI Law**, which would be "a coordinator for existing efforts to govern AI and catalyze multilateral treaties and arrangements for neglected issues."[99]

3.5 Critiques of this model: Castel and Castel have critiqued international conventions on the grounds that they "are difficult to monitor and control."[100] More specifically, Ho and others have argued that a model like an Advanced AI Governance Organization would face challenges around its ability to set and update standards sufficiently quickly; around incentivizing state participation in adopting the regulations, and in sufficiently scoping the challenges to focus on.[101] Finally, reviewing general patterns in current state activities on AI standard-setting, von Ingersleben has noted that "technical experts hailing from geopolitical rivals, such as the United States and China, readily collaborate on technical AI standards within transnational standard-setting organizations, whereas governments are much less willing to collaborate on global ethical AI standards within international organizations,"[102] which suggests potential thresholds to overcoming State disinterest in participating in any international institutions focused on more political and ethical standard-setting.

Model 4: Enforcement of standards or restrictions

4.1 Functions and types: The function of this institutional model is to prevent the production, proliferation or irresponsible deployment of a dangerous or illegal technology, product or activity. To fulfill that function, institutions under this model rely, among other mechanisms, on (1) bans and moratoria, (2) non-proliferation regimes, (3) export control lists, (4) monitoring and verification mechanisms,[103] (5) licensing regimes, and (6) registering and/or tracking of key resources, materials, or stocks, amongst others. Other types of mechanisms, such as (7) confidence-building measures

[99] Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[100] Castel, J.G., and Mathew E. Castel. 'The Road to Artificial Superintelligence - Has International Law a Role to Play?' Canadian Journal of Law & Technology 14 (2016). https://ojs.library.dal.ca/CJLT/article/download/7211/6256.; pg 11

[101] Ho and others, 'International Institutions for Advanced AI', p. 10-11.

[102] Ingersleben-Seip, Nora von. 'Competition and Cooperation in Artificial Intelligence Standard Setting: Explaining Emergent Patterns'. Review of Policy Research 40, no. 5 (25 January 2023): 781–810. https://doi.org/10.1111/ropr.12538.

[103] Monitoring and verification arrangements can come in a range of forms. For instance, some institutional agreements enable bilateral 'open monitoring' (e.g. enable intrusive inspections); others provide for 'closed monitoring' (e.g. unilateral monitoring through spy satellites or plane overflight). For the distinction, see also: Coe, Andrew J., and Jane Vaynman. 'Why Arms Control Is So Rare'. *American Political Science Review* 114, no. 2 (May 2020): 342–55. https://doi.org/10.1017/S000305541900073X.

(CBMs), are generally transparency-enabling.[104] While generally focused on managing tensions and preventing escalations,[105] CBMs can also build trust amongst States in each others' mutual compliance with standards or prohibitions, and can therefore also support or underwrite standards- and restriction-enforcing institutions.

4.2 Common examples: The most prominent example of this model, especially in discussions of institutions capable of carrying out monitoring and verification roles, is the International Atomic Energy Agency (IAEA)[106]—in particular, its Department of Safeguards. Many other proposals refer to the monitoring & verification mechanisms of arms control treaties.[107] For instance, Baker has studied the monitoring and verification mechanisms for different types of nuclear arms control regimes, reviewing the role of the IAEA system under Comprehensive Safeguards Agreements with Additional Protocols in monitoring nonproliferation treaties such as the Non-Proliferation Treaty (NPT) and the five Nuclear-Weapon-Free-Zone Treaties; the role of monitoring and verification arrangements in monitoring of bilateral nuclear arms control limitation treaties, and

[104] For definitions of CBMs, see: Horowitz, Michael C, and Paul Scharre. 'AI and International Stability: Risks and Confidence-Building Measures'. Center for a New American Security, 12 January 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures. Pg. 4. (–'unilateral, bilateral, and/or multilateral actions that states can take to build trust and prevent inadvertent military conflict. [...] generally involve using transparency, notification, and monitoring to attempt to mitigate the risk of conflict.'). For another definition see also: Horowitz, Michael C., Lauren Kahn, and Casey Mahoney. 'The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?' *Orbis*, 14 September 2020. https://doi.org/10.1016/j.orbis.2020.08.003. ('a class of information-sharing and transparency-enhancing arrangements').

[105] In line with the stabilization model, discussed in the next section.

[106] For invocations of the IAEA and NPT examples to AI governance, see also: 'Secretary-General António Guterres remarks to the Security Council on Artificial Intelligence'; https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence; UN Press. 'Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence', 18 July 2023. https://press.un.org/en/2023/sgsm21880.doc.htm.; Robinson, Mary. 'The Elders Urge Global Co-Operation to Manage Risks and Share Benefits of AI', 31 May 2023. https://theelders.org/news/elders-urge-global-co-operation-manage-risks-and-share-benefits-ai.; Altman, Sam, Greg Brockman, and Ilya Sutskever. 'Governance of Superintelligence'. OpenAI, 22 May 2023. https://openai.com/blog/governance-of-superintelligence. Ramamoorthy, Anand, and Roman Yampolskiy. 'Beyond MAD?: The Race for Artificial General Intelligence'. *ITU JOURNAL: ICT DISCOVERIES* 1, no. 1 (2018): 8. https://www.itu.int/en/journal/001/Documents/itu2018-9.pdf ; see also Chesterman, Simon. *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law*. Cambridge: Cambridge University Press, 2021. https://doi.org/10.1017/9781009047081. Pg. 210. The Nuclear Non-Proliferation Regime (NPT) is also discussed by Robichaud, Carl. 'The Puzzle of Non-Proliferation'. *Asterisk*, June 2023. https://asteriskmag.com/issues/03/the-puzzle-of-non-proliferation; and in Maas, Matthijs M. 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons'. *Contemporary Security Policy* 40, no. 3 (6 February 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.

[107] Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims', 20 April 2020. http://arxiv.org/abs/2004.07213. pg . 67-69.

the role of the International Monitoring System (IMS) in monitoring and enforcing nuclear test bans.[108] Shavit has likewise referred to the precedent of the NPT and IAEA in discussing a resource (compute) monitoring framework for AI.[109] Gutierrez has invoked Interpol's 'red notice' alert system as an example of a model by which an international institution could alert global stakeholders about the dangers of a particular AI system.[110]

Examples given of export control regimes include the Nuclear Suppliers Group, the Wassenaar Arrangement and the Missile Technology Control Regime.[111] As examples of CBMs, people have pointed to the Open Skies Treaty,[112] which is enforced by the Open Skies Consultative Commission.

There are also examples of global technology control institutions that were not carried through, but which are still discussed as inspirations for AI, such as the international Atomic Development Authority (ADA) proposed in the early nuclear age;[113] or early- to mid-20th-century proposals for the global regulation of military aviation.[114]

4.3 Underexplored examples: Examples that are not yet often discussed but that could be promising are the Organisation for the Prohibition of Chemical Weapons (OPCW),[115] the Biological Weapons Convention's Implementation Support Unit, the International Maritime Organization (in its ship registration function) and the Convention on International Trade in Endangered Species of Wild Fauna and Flora's (CITES) Secretariat, specifically, its database of national import and export reports.

[108] Baker, Mauricio. 'Nuclear Arms Control Verification and Lessons for AI Treaties'. arXiv, 8 April 2023. http://arxiv.org/abs/2304.04123.

[109] Shavit, Yonadav. 'What Does It Take to Catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training Via Compute Monitoring', 2023. https://paperswithcode.com/paper/what-does-it-take-to-catch-a-chinchilla. Pg. 2.

[110] Gutierrez, Carlos I. 'Multilateral Coordination for the Proactive Governance of Artificial Intelligence Systems'. Future of Life Institute, forthcoming 2023.

[111] Maas 'Artificial Intelligence Governance Under Change: Foundations, Facets, Frameworks'. Ftn 242.

[112] Hobbhahn, Marius, Max Räuker, Yannick Mühlhäuser, Jasper Götting, and Simon Grimm. 'What Success Looks Like'. Effective Altruism Forum, 28 June 2022. https://forum.effectivealtruism.org/posts/AuRBKFnjABa6c6GzC/what-success-looks-like.

[113] Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In *Risks of Artificial Intelligence*. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey.

[114] Zaidi, Waqar H., ed. 'Conclusion: Science, Technology, and Internationalism into the Cold War and Beyond'. In *Technological Internationalism and World Order*, 239–47. Science in History. Cambridge: Cambridge University Press, 2021. https://doi.org/10.1017/9781108872416.009.

[115] See also Whitfield, Robert. 'Effective, Timely and Global: The Urgent Need for Good Global Governance of AI'. World Federalist Movement and Institute for Global Policy, 2020. https://www.wfm-igp.org/publication/effective-timely-and-global-the-urgent-need-for-good-global-governance-of-ai/. pg. 63.

4.4 Proposed AI institutions along this model: Proposals along this model are particularly widespread and prevalent. Indeed, as mentioned, a significant part of the literature on the international governance of AI has made reference to some sort of 'IAEA for AI'. For instance, in relatively early proposals,[116] Turchin and others propose a '**UN-backed AI-control agency**' which "would require much tighter and swifter control mechanisms, and would be functionally equivalent to a world government designed specifically to contain AI."[117] Ramamoorthy and Yampolskiy have proposed a '**global watchdog agency**' that would have the express purpose of tracking AGI programs, and that would have the jurisdiction and the lawful authority to intercept and halt unlawful attempts at AGI development.[118] Pointing to the precedent of both the IAEA and its inspection regime, and the Comprehensive Nuclear Test-Ban Treaty Organization (CTBTO)'s Preparatory Commission, Nindler has proposed an **International Enforcement Agency** for safe AI research and development, which would support and implement the provisions of an international treaty on safe AI research and development, with the general mission "to accelerate and enlarge the contribution of artificial intelligence to peace, health and prosperity throughout the world [and … to ensure that its assistance] is not used in such a way as to further any military purpose."[119] Such a body would be charged with drafting safety protocols and measures, and he suggests that its enforcement could, in extreme cases, be backed up by the use of force under the UN Security Council's use of its Chapter VII powers.[120]

Whitfield has drawn on the example of the UN Framework Convention on Climate Change to propose a UN Framework Convention on AI (UNFCAI) along with a Protocol on AI that would subsequently deliver the first set of enforceable AI regulations. He proposes that these should be supported by three new bodies: an **AI Global Authority (AIGA)** to provide an inspection regime in particular for military AI, an associated **'Parliamentary Assembly' supervisory body** that would enhance democratic input into the treaty's operations and play 'a constructive monitoring role', as well as a

[116] For an older review of even earlier proposals, some of which envisioned global regulation and/or monitoring and enforcement, see: Sotala, Kaj, and Roman V Yampolskiy. 'Responses to Catastrophic AGI Risk: A Survey'. Physica Scripta 90, no. 1 (1 January 2015): 018001. https://doi.org/10.1088/0031-8949/90/1/018001. And Sotala, Kaj, and Roman V. Yampolskiy. 'Responses to Catastrophic AGI Risk: A Survey.' Technical Report. Berkeley, CA: Machine Intelligence Research Institute, 2013. https://intelligence.org/files/ResponsesAGIRisk.pdf.

[117] Turchin, Alexey, David Denkenberger, and Brian Patrick Green. 'Global Solutions vs. Local Solutions for the AI Safety Problem'. *Big Data and Cognitive Computing* 3, no. 1 (March 2019): 16. https://doi.org/10.3390/bdcc3010016. Pg. 4.

[118] Ramamoorthy, Anand, and Roman Yampolskiy. 'Beyond MAD?: The Race for Artificial General Intelligence'. ITU JOURNAL: ICT DISCOVERIES 1, no. 1 (2018): 8. https://www.itu.int/en/journal/001/Documents/itu2018-9.pdf

[119] Nindler, Reinmar. 'The United Nation's Capability to Manage Existential Risks with a Focus on Artificial Intelligence'. International Community Law Review 21, no. 1 (11 March 2019): 5–34. https://doi.org/10.1163/18719732-12341388. Pg. 31.

[120] Ibid. 32.

multistakeholder **Intergovernmental Panel on AI**, to provide scientific, technical and policy advice to the UNFCAI.[121]

More recently,[122] Ho and others have proposed an '**Advanced AI Governance Organization**' which, in addition to setting international standards for the development of advanced AI (as discussed above), could monitor compliance with these standards through, for example, self-reporting, monitoring practices within jurisdictions or detection and inspection of large data centers.[123] Altman and others have proposed an '**AIEA for Superintelligence**' consisting of "an international authority that can inspect systems, require audits, test for compliance with safety standards, place restrictions on degrees of deployment and levels of security."[124] In a very similar vein, Guest (based on an earlier proposal by Karnofsky)[125] has called for an '**International Agency for Artificial Intelligence (IAIA)**' to conduct "extensive verification through on-chip mechanisms [and] on-site inspections, as part of his proposal for a 'Collaborative Handling of Artificial Intelligence Risks with Training Standards (CHARTS)'.[126] Drawing together elements from several models—and referring to the examples of the IPCC, Interpol, and the WTO's dispute settlement system—Gutierrez has proposed a '**multilateral AI governance initiative**' to mitigate 'the shared large-scale high-risk harms caused directly or indirectly by AI'.[127] His proposal envisions an organizational structure consisting of (1) a forum for member state representation (which adopts decisions via supermajority); (2) technical bodies, such as an external board of experts, and a permanent technical and liaison secretariat that works from an information and enforcement network, and which can issue 'red notice' alerts; and (3) an arbitration board that can hear both complaints by non-state AI developers who seek to contest these notices, as well as member states.[128]

Previously, Wilson has proposed an '**Emerging Technologies Treaty**'[129] that would address risks from many emerging technologies, and which in his view could either be housed under an existing international organization or body, or established separately,

[121] Whitfield, 'Effective Timely and Global: The Urgent Need for Good Global Governance of AI'.

[122] As discussed in Model 3, above.

[123] Ho and others, 'International Institutions for Advanced AI', p. 9-10.

[124] Altman, Sam, Greg Brockman, and Ilya Sutskever. 'Governance of Superintelligence'.

[125] Karnofsky, Holden. 'Nearcast-Based "Deployment Problem" Analysis'. LessWrong 2.0, 21 September 2022. https://www.lesswrong.com/posts/vZzg8NS7wBtqcwhoJ/nearcast-based-deployment-problem-analysis (see quote).

[126] Oliver Guest, 'Prospects for AI safety agreements between countries' (Rethink Priorities, 2023) https://rethinkpriorities.org/publications/prospects-for-ai-safety-agreements-between-countries

[127] Gutierrez, 'Multilateral Coordination for the Proactive Governance of Artificial Intelligence Systems'.

[128] ibid.

[129] Wilson, Grant. 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law'. Va. Envtl. LJ 31 (2013): 307. http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/velj31&section=12

and which would establish a body of experts that would determine whether there was a 'reasonable grounds for concern' about AI or other dangerous research, after which States would be required to regulate or temporarily prohibit research.[130] Likewise drawing on the IAEA model, Chesterman has proposed an **International Artificial Intelligence Agency (IAIA)** as an institution with "a clear and limited normative agenda, with a graduated approach to enforcement", arguing that "the main 'red line' proposed here would be the weaponization of AI—understood narrowly as the development of lethal autonomous weapon systems lacking 'meaningful human control' and more broadly as the development of AI systems posing a real risk of being uncontrollable or uncontainable".[131] In practice, this organization would draw up safety standards, develop a forensic capability to identify those responsible for 'rogue' AI, serve as a clearinghouse to gather and share information about such systems, and to provide early notification of emergencies.[132] Chesterman argues that one organizational cause that could be adopted for this IAIA, is to learn from the IAEA, where its Board of Governors (rather than the annual General Conference) has ongoing oversight of its operations.

Other authors endorse an institution more directly aimed at preventing or limiting proliferation of dangerous AI systems. Jordan and others have proposed a '**NPT+**' model;[133] the Future of Life Institute (FLI) has proposed '**international agreements** to limit particularly high-risk AI proliferation and mitigate the risks of advanced AI'.[134] PauseAI has proposed an international agreement that sets up an '**International AI Safety Agency**' that would be in charge of granting approvals for deployments of AI systems and new training runs above a certain size.[135] The Elders, a group of independent former world leaders, have recently called on countries to request, via the UN General Assembly, that the International Law Commission draft an international treaty to establish a new '**International AI Safety Agency**',[136] drawing on the models of the NPT and the IAEA, "to manage these powerful technologies within robust safety protocols [and to ...] ensure AI is used in ways consistent with international law and

[130] Ibid. pg. 345-355.

[131] Chesterman, Simon. We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law. Cambridge: Cambridge University Press, 2021. https://doi.org/10.1017/9781009047081. Pg. 209-217. And previously Chesterman, Simon. 'Weapons of Mass Disruption: Artificial Intelligence and International Law'. Cambridge International Law Journal, 23 April 2021. https://papers.ssrn.com/abstract=3832563.

[132] Ibid. pg. 216.

[133] Jordan, Richard, Nicholas Emery-Xu, and Robert Trager. 'International Governance of Artificial Intelligence', (working paper).

[134] FLI. 'FLI on "A Statement on AI Risk" and Next Steps'. *Future of Life Institute* (blog), 30 May 2023. https://futureoflife.org/ai-policy/fli-on-a-statement-on-ai-risk-and-next-steps/.

[135] PauseAI. 'PauseAI Proposal'. Accessed 28 August 2023. https://pauseai.info/proposal.

[136] Robinson, Mary. 'The Elders Urge Global Co-Operation to Manage Risks and Share Benefits of AI', 31 May 2023. https://theelders.org/news/elders-urge-global-co-operation-manage-risks-and-share-benefits-ai.

human rights treaties".[137] More specific monitoring provisions are also entertained; for instance, Balwit briefly discusses an **advanced AI chips registry**, potentially organized by an international agency.[138]

At the level of transparency-supporting agreements, there are many proposals for Confidence-Building Measures for (military) AI. Such proposals focus on bilateral arrangements that build confidence amongst States and contribute to stability (as under Model 5), but which lack distinct institutions. For instance, Shoker and others discuss an '**international code of conduct for state behavior**'.[139] Scharre, Horowitz, Khan and others have discussed a range of other AI CBMs;[140] including the marking of autonomous weapons systems, geographic limits, limits on particular (e.g. nuclear) operations of AI.[141] They propose to group these under a **International Autonomous Incidents Agreement (IAIA)** to "help reduce risks from accidental escalation by autonomous systems, as well as reduce ambiguity about the extent of human intention behind the behavior of autonomous systems."[142] In doing so, they have pointed to the precedent of arrangements such as the 1972 Incidents at Sea Agreement,[143] as well as the 12th-19th century development of Maritime Prize Law.[144] Imbrie and Kania have

[137] Ibid.

[138] Balwit, Avital. 'How We Can Regulate AI'. *Asterisk*, June 2023. https://asteriskmag.com/issues/03/how-we-can-regulate-ai.

[139] Shoker, Sarah, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, et al. 'Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings'. arXiv, 3 August 2023. https://doi.org/10.48550/arXiv.2308.00862.

[140] Ibid. See also Ruhl, Christian. 'Autonomous Weapon Systems & Military AI: Cause Area Report'. Founders Pledge, May 2022. https://founderspledge.com/stories/autonomous-weapon-systems-and-military-artificial-intelligence-ai.; Horowitz, Michael C., and Lauren Kahn. 'How Joe Biden Can Use Confidence-Building Measures for Military Uses of AI'. *Bulletin of the Atomic Scientists* 77, no. 1 (2 January 2021): 33–35. https://doi.org/10.1080/00963402.2020.1860331.; Horowitz, Michael C., Lauren Kahn, and Casey Mahoney. 'The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?' *Orbis*, 14 September 2020. https://doi.org/10.1016/j.orbis.2020.08.003.

[141] Horowitz, Michael C, and Paul Scharre. 'AI and International Stability: Risks and Confidence-Building Measures'. Center for a New American Security, 12 January 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[142] Horowitz, Michael C, and Paul Scharre. 'AI and International Stability: Risks and Confidence-Building Measures'. Center for a New American Security, 12 January 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures , pg 11.

[143] Ruhl, Christian. 'Autonomous Weapon Systems & Military AI: Cause Area Report'. Founders Pledge, May 2022. https://founderspledge.com/stories/autonomous-weapon-systems-and-military-artificial-intelligence-ai., pg. 39-40; Horowitz, Michael C., and Lauren Kahn. 'How Joe Biden Can Use Confidence-Building Measures for Military Uses of AI'. Bulletin of the Atomic Scientists 77, no. 1 (2 January 2021): 33–35. https://doi.org/10.1080/00963402.2020.1860331.

[144] Ibid.

proposed an '**Open Skies on AI**' agreement.[145] Bremmer & Suleyman have proposed a **bilateral US-China regime** to foster cooperation between the US and Beijing on AI, envisioning this 'to create areas of commonality and even guardrails proposed and policed by a third party.'[146]

4.5 Critiques of this model: Many critiques of the enforcement model have ended up focusing (whether fairly or not) on the appropriateness of the basic analogy between nuclear weapons and AI, that is explicit or implicit in proposals for an 'IAEA' or 'NPT'-like regime. For instance, Kaushik and Korda have opposed what they see as aspirations to a 'wholesale ban' on dangerous AI, and argue that 'attempting to regulate artificial intelligence indiscriminately would be akin to regulating the concept of nuclear fission itself.'[147]

Others critique the appropriateness of an IAEA-modeled approach: Stewart suggests that the focus on the IAEA's safeguards is inadequate since AI systems cannot be safeguarded in the same way; and suggests that rather, better lessons might be found in the IAEA's International Physical Protection Advisory Service (IPPAS) missions, which allow it to serve as an independent third party to assess the regulatory preparedness of countries that aim to develop nuclear programs.[148] Drexel and Depp have argued that even if this IAEA model could work on a technical level, it will likely be prohibitively difficult to negotiate such an intense level of oversight.[149] Further, Sepasspour as well as Law have noted that rather than a straightforward set-up, there were years of delay between the IAEA's establishment (1957), its adoption of the INFCIRC 26 safeguards document (1961), its taking of a leading role in nuclear nonproliferation upon the adoption of the NPT (1968), and its eventual further empowerment of its verification

[145] Imbrie, Andrew, and Elsa B. Kania. 'AI Safety, Security, and Stability Among Great Powers: Options, Challenges, and Lessons Learned for Pragmatic Engagement'. CSET Policy BRief. Center for Security and Emerging Technology, December 2019. https://cset.georgetown.edu/wp-content/uploads/AI-Safety-Security-and-Stability-Among-the-Great-Powers.pdf.

[146] Bremmer, Ian, and Mustafa Suleyman. 'The AI Power Paradox'. *Foreign Affairs*, 16 August 2023. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox.

[147] See Kaushik, Divyansh, and Matt Korda. 'Panic about Overhyped AI Risk Could Lead to the Wrong Kind of Regulation'. Vox, 3 July 2023. https://www.vox.com/future-perfect/2023/7/3/23779794/artificial-intelligence-regulation-ai-risk-congress-sam-altman-chatgpt-openai. See also Watson, Mike. 'IAEA for AI? That Model Has Already Failed'. *Wall Street Journal*, 1 June 2023, sec. Opinion. https://www.wsj.com/articles/iaea-for-ai-that-model-has-already-failed-chaptgpt-technology-nuclear-proliferation-4339543b.

[148] Stewart, John. 'Why the IAEA Model May Not Be Best for Regulating Artificial Intelligence'. Bulletin of the Atomic Scientists (blog), 9 June 2023. https://thebulletin.org/2023/06/why-the-iaea-model-may-not-be-best-for-regulating-artificial-intelligence/.

[149] Drexel, Bill Drexel, Michael, and Michael Depp. 'Every Country Is on Its Own on AI: Why AI Regulation Can't Follow in the Footsteps of International Nuclear Controls.' Foreign Policy (blog), 13 June 2023. https://foreignpolicy.com/2023/06/13/ai-regulation-international-nuclear/.

function through the Additional Protocol (1997).[150] Such a slow aggregation might not be adequate, given the speed of advanced AI development. Finally, another issue is that the strength of an IAEA agency depends on the existence of supportive international treaties, as well as specific incentives for participation.

Others question whether this model would even be desirable, even if achievable. Howard has generally critiqued many governance proposals that would involve centralized control (whether domestic or global) over the proliferation of- and access to frontier AI systems, arguing that such centralisation would end up only advantaging currently powerful AI labs as well as malicious actors willing to steal models, with the concern that this would have significant illiberal effects.[151]

Model 5: Stabilization and emergency response

5.1 Functions and types: The function of this institutional model is to ensure that an emerging technology or an emergency does not have a negative impact on social stability and international peace.

Such institutions can serve various subsidiary functions, including (1) *general stability management*, by assessing and mitigating systemic vulnerabilities that are susceptible to incidents or accidents; (2) provide *early warning* of—and *response coordination* to—incidents and emergencies, providing timely warning, and creating common knowledge of an emergency;[152] (3) generally *stabilizing relations, behavior and expectations* around AI technology to encourage transparency and trust around State activities in a particular domain, and to avoid inadvertent military conflict.

5.2 Common examples: Examples of institutions involved in stability management include the Financial Stability Board (FSB), an entity 'composed of central bankers, ministries of finance, and supervisory and regulatory authorities from around the world'.[153] Another example might be the United Nations Office for Disaster Risk Reduction (UNDRR), which focuses on responses to natural disasters.[154]

[150] Law, Harry. 'An IAEA for AI? The Early History of the International Atomic Energy Agency'. Harry Law, 9 June 2023. https://www.harrylaw.co.uk/post/an-iaea-for-ai-the-early-history-of-the-international-atomic-energy-agency. Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. Bulletin of the Atomic Scientists, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249. Pg. 312.

[151] Howard, Jeremy. 'AI Safety and the Age of Disenlightenment'. fast.ai, 10 July 2023. https://www.fast.ai/posts/2023-11-07-dislightenment.html.

[152] Yudkowsky, Eliezer. 'There's No Fire Alarm for Artificial General Intelligence'. *Machine Intelligence Research Institute* (blog), 14 October 2017. https://intelligence.org/2017/10/13/fire-alarm/.

[153] Bremmer, Ian, and Mustafa Suleyman. 'The AI Power Paradox'. *Foreign Affairs*, 16 August 2023. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradox

[154] See also Whitfield, 'Effective Timely and Global: The Urgent Need for Good Global Governance of AI'.

5.3 Underexplored examples: Examples that are not yet invoked, but that could be promising examples of early warning functions include WHO's 'public health emergency of international concern' early warning mechanism, or the procedure established in the IAEA's 1986 Convention on Early Notification of a Nuclear Accident.

5.4 Proposed AI institutions along this model: AI proposals along the early warning model include Pauwels' proposal for a **Global Foresight Observatory**, as a multi-stakeholder platform aimed at fostering greater cooperation in technological and political preparedness for the impacts of innovation in various fields, including AI.[155] Brenner and Suleyman's proposal for a **Geotechnology Stability Board** which "could work to maintain geopolitical stability amid rapid AI-driven change" based on the coordination of national regulatory authorities and international standard-setting bodies. At other times, such a body would help prevent global technology actors from "engaging in regulatory arbitrage or hiding behind corporate domiciles." Finally, it could also take up responsibility for governing open-source AI, and censoring the uploading of highly dangerous models.[156]

5.5 Critiques of this model: As there have been relatively limited numbers of proposals for this model, there are not yet many critiques. However, possible critiques might focus on the potential adequacy of relying on international institutions to respond to (rather than prevent) situations where dangerous AI systems have already seen deployment, as in those situations coordinating, communicating and implementing effective countermeasures might either be very difficult, or far too slow to respond adequately to countering a misaligned AI system.

Model 6: International joint research

6.1 Functions and types: The function of this institutional model is to start a bilateral or multilateral collaboration between States or State entities to solve a common problem or achieve a common goal. Most of these models would focus on accelerating the development of a technology or the exploitment of a resource by particular actors to avoid races. Other models would aim at speeding up the development of safety techniques.

In some proposals, an institution like this aims not just to rally and organize a major research project, but simultaneously aims to include elements of an enforcing institution, in order to exclude all other actors from conducting research and/or creating capabilities around that problem or goal, creating a de facto or an explicit international monopoly on an activity.

[155] Pauwels, Eleonore. The new geopolitics of converging risks: the UN and prevention in the era of AI, Centre for Policy Research, UNU, 2 May 2019, 53, https://cpr.unu.edu/the-new-geopolitics-of-converging-risksthe-un-and-prevention-in-the-era-of-ai.html.

[156] Bremmer, Ian, and Mustafa Suleyman. 'The AI Power Paradox'. *Foreign Affairs*, 16 August 2023. https://www.foreignaffairs.com/world/artificial-intelligence-power-paradoxIbid.

6.2 Common examples: Examples that are pointed to as models of an international joint scientific program include the European Organization for Nuclear Research (CERN),[157] ITER, the International Space Station (ISS), or the Human Genome Project.[158] Examples of models of a (proposed) international monopoly include the Acheson-Lilienthal Proposal,[159] and the resulting Baruch Plan, which called for the creation of an Atomic Development Authority.[160]

6.3 Underexplored examples: Examples that are not yet discussed but that could be promising are the James Webb Space Telescope, and the Laser Interferometer Gravitational-Wave Observatory (LIGO),[161] which is organized internationally through the LIGO Scientific Collaboration (LSC).

6.4 Proposed AI institutions along this model: Explicit AI proposals along the joint scientific program model are various.[162] Some proposals focus primarily on accelerating

[157] Marcus, Gary. 'Artificial Intelligence Is Stuck. Here's How to Move It Forward.' *The New York Times*, 29 July 2017, sec. Opinion. https://www.nytimes.com/2017/07/29/opinion/sunday/artificial-intelligence-is-stuck-heres-how-to-move-it-forward.html.; Marcus, Gary. 'Two Models of AI Oversight — and How Things Could Go Deeply Wrong'. *Communications of the ACM*, 12 June 2023. https://cacm.acm.org/blogs/blog-cacm/273791-two-models-of-ai-oversight-and-how-things-could-go-deeply-wrong/fulltext; Miotti, Andrea. 'We Can Prevent AI Disaster Like We Prevented Nuclear Catastrophe'. *Time*, 15 September 2023. https://time.com/6314045/prevent-ai-disaster-nuclear-catastrophe/. For a draft study on the lessons from CERN for international institutions for AI, see also Frazier, Kevin, 'CERN case study', 2023. https://docs.google.com/document/d/1_-VtICdXJPYgQwUF5UGPG_xRAs3zXWrVpF8v8mE8fPI/edit#heading=h.kepzessvh3h6

[158] Levin, John-Clark, and Matthijs M. Maas. 'Roadmap to a Roadmap: How Could We Tell When AGI Is a "Manhattan Project" Away?', 7. Santiago de Compostela, Spain, 2020. http://dmip.webs.upv.es/EPAI2020/papers/EPAI_2020_paper_11.pdf; Castel, J.G., and Mathew E. Castel. 'The Road to Artificial Superintelligence - Has International Law a Role to Play?' *Canadian Journal of Law & Technology* 14 (2016). https://ojs.library.dal.ca/CJLT/article/download/7211/6256. Pg. 11.

[159] Jordan, Richard, Nicholas Emery-Xu, and Robert Trager. 'International Governance of Artificial Intelligence', (working paper).

[160] Zaidi, Waqar, and Allan Dafoe. 'International Control of Powerful Technology: Lessons from the Baruch Plan'. Center for the Governance of AI, Future of Humanity Institute, March 2021. https://www.fhi.ox.ac.uk/wp-content/uploads/2021/03/International-Control-of-Powerful-Technology-Lessons-from-the-Baruch-Plan-Zaidi-Dafoe-2021.pdf. See also Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In Risks of Artificial Intelligence. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. pg. 7.

[161] Levin and Maas, 'Roadmap to a Roadmap: How Could We Tell When AGI Is a "Manhattan Project" Away?'; See generally: Robinson, Mark. 'Big Science Collaborations; Lessons for Global Governance and Leadership'. *Global Policy* 12(1) 66-80 (2020). https://doi.org/10.1111/1758-5899.12861.

[162] Note, these proposals are distinct from (past) calls for individual States to accelerate AI research and/or undertake some sort of large-scale AI research project or sprint. See for example: Hammond, Samuel. 'We Need a Manhattan Project for AI Safety'. POLITICO, 8 May 2023. https://www.politico.com/news/magazine/2023/05/08/manhattan-project-for-ai-safety-00095779. ;

safety. Lewis Ho and others suggest an '**AI Safety Project**' to "promote AI safety R&D by promoting its scale, resourcing and coordination". To ensure AI systems are reliable and less vulnerable to misuse, this institution would have access to significant compute and engineering capacity, as well as AI models developed by AI companies. Contrary to other international joint scientific programs like CERN or ITER, which are strictly inter-governmental, Ho and others propose that the AI Safety Project comprise other actors as well (e.g. civil society and the industry). The authors also suggest that, to prevent replication of models or diffusion of dangerous technologies, the AI Safety Project should incorporate information and security measures such as siloing information, structuring model access and designing internal review processes.[163] Neufville and Baum have pointed out that 'a **clearinghouse for research into AI**' could solve the collective problem of underinvestment in basic research, AI ethics and safety research.[164] More ambitiously, Ramamoorthy and Yampolskiy have previously proposed a '**Benevolent AGI Treaty**', which involves 'the development of AGI as a global, non-strategic humanitarian objective, under the aegis of a special agency within the United Nations'.[165]

Other proposals suggest inter-governmental collaboration for the development of AI systems more generally. Daniel Zhang and others at Stanford University's HAI recommend the United States and like-minded allies create a '**Multilateral Artificial Intelligence Research Institute (MAIRI)**' to facilitate scientific exchanges and promote collaboration on AI research—including the risks, governance, and socio-economic impact of AI—based on a foundational agreement outlining agreed research practices. The authors suggest that MAIRI could also strengthen policy coordination around AI.[166] Fischer and Wenger add that a '**neutral hub for AI research**' should have four functions: (i) fundamental research in the field of AI; (ii) research and reflection on societal risks associated with AI; (iii) development of norms and best practices regarding the application of AI; and (iv) further education for AI researchers. This hub could be created by a conglomerate of like-minded States, but should eventually be open to all States and possibly be linked to the United Nations

or previously McGinnis, John O. 'Accelerating AI'. Northwestern University Law Review 104 (2010). https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1193&context=nulr_online.

[163] Ho and others, 'International Institutions for Advanced AI', p 13-14.

[164] Neufville, Robert de, and Seth D. Baum. 'Collective Action on Artificial Intelligence: A Primer and Review'. Technology in Society 66 (1 August 2021): 101649. https://doi.org/10.1016/j.techsoc.2021.101649.

[165] Ramamoorthy, Anand, and Roman Yampolskiy. 'Beyond MAD?: The Race for Artificial General Intelligence'. ITU JOURNAL: ICT DISCOVERIES 1, no. 1 (2018): 8. https://www.itu.int/en/journal/001/Documents/itu2018-9.pdf pg 5.

[166] Daniel Zhang and others, 'Enhancing International Cooperation in AI Research: The Case for a Multilateral AI Research Institute' (HAI, 2022) https://hai.stanford.edu/white-paper-enhancing-international-cooperation-ai-research-case-multilateral-ai-research-institute

through a cooperation agreement, according to the authors.[167] Other authors posit that an international collaboration on AI research and development should include all members of the United Nations from the start, as similar projects like the ISS or the Human Genome Project have done. They suggest this approach might reduce the possibility of an international conflict.[168] In this vein, Kemp and others call for the foundation of a '**UN AI Research Organization (UNAIRO)**', which would focus on 'building AI technologies in the public interest, including to help meet international targets [...] [a] secondary goal could be to conduct basic research on improving AI techniques in the safest, careful and responsible environment possible.'[169]

Philipp Slusallek, Scientific Director of the German Research Center for Artificial Intelligence, suggests a '**CERN for AI**'—"a collaborative, scientific effort to accelerate and consolidate the development and uptake of AI for the benefit of all humans and our environment." Slusallek promotes a very open and transparent design for this institution, in which data and knowledge would flow freely between collaborators.[170] Similarly, the Large-scale Artificial Intelligence Open Network (LAION) calls for a CERN-like open source collaboration among the United States and allied countries to establish an **international "supercomputing research facility"** hosting "a diverse array of machines equipped with at least 100,000 high-performance state-of-the-art accelerators" that can be overseen by democratically elected institutions from participating countries.[171] Daniel Dewey goes a step further and suggests a potential **joint international AI project with a monopoly** over hazardous AI development, in the same spirit of the 1946 Baruch Plan, which proposed an International Atomic Development Authority with a monopoly over nuclear activities. However, Dewey admits this proposal is possibly politically intractable.[172] In another proposal for monopolized

[167] Fischer, Sophie-Charlotte, and Andreas Wenger. 'A Politically Neutral Hub for Basic AI Research'. Policy Perspectives. Zurich: CSS, ETH Zurich, March 2019. http://www.css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/PP7-2_2019-E.pdf.

[168] Castel, J.G., and Mathew E. Castel. 'The Road to Artificial Superintelligence - Has International Law a Role to Play?' Canadian Journal of Law & Technology 14 (2016). https://ojs.library.dal.ca/CJLT/article/download/7211/6256. (pg 11-12).

[169] Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[170] Slusallek, P., Artificial Intelligence and Digital Reality: Do We Need a CERN for AI? the OECD Forum Network, 2018. January 8, https://www.oecd-forum.org/posts/28452-artificial-intelligence-and-digital-reality-do-we-need-a-cern-for-ai.

[171] Schuhmann, Christoph, 'Petition for keeping up the progress tempo on AI research while securing its transparency and safety' (LAION, 29 March 2023), https://laion.ai/blog/petition/

[172] Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In Risks of Artificial Intelligence. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. Pg. 7.

international development, Miotti has proposed a '**Multilateral AGI Consortium**' (MAGIC), which would be an international organization mandated to run 'the world's only advanced and secure AI facility focused on safety-first research and development of advanced AI.'[173] This organization would only share breakthroughs with the outside world once proven demonstrably safe, and would therefore be coupled with a global moratorium on the creation of AI systems exceeding a set compute governance threshold.

The proposals for an institution analogous to CERN discussed thus far envision a grand institution that draws talent and resources for research and development of AI projects in general. Other proposals have a more narrow focus. Charlotte Stix, for example, suggests that a more decentralized version of this model could be more beneficial. Stix argues that a '**European Artificial Intelligence megaproject**' could be composed of a centralized headquarters to overview collaborations and provide economies of scale for AI precursors within a network of affiliated AI laboratories that conduct most of the research.[174] Other authors argue that rather than focus on AI research in general, an international research collaboration could focus on the use of AI to solve problems in a specific field, such as climate change, health, privacy-enhancing technologies, economic measurement, or the sustainable development goals.[175]

6.5 Critiques of this model: In general, there have been few sustained critiques of this institutional model. However, Ho and others have suggested that an international collaboration to conduct technical AI safety research might face challenges, in that it might pull safety researchers away from the frontier AI developers, reducing the amount of in-house safety expertise. In addition, there are concerns that any international project that would need to access advanced AI models, would run risks over security concerns and model leaking.[176]

Moreover, more fundamental critiques do exist: for instance, Kaushik and Korda have critiqued the feasibility of a 'Manhattan Project-like undertaking to address the alignment problem', arguing that massively accelerating AI safety research through any

[173] Miotti, 'We Can Prevent AI Disaster Like We Prevented Nuclear Catastrophe'. This is similar to a proposal by Hogarth, Ian. 'We Must Slow down the Race to God-like AI'. Financial Times, 13 April 2023. https://www.ft.com/content/03895dc4-a3b7-481e-95cc-336a524f2ac2.

[174] Stix, Charlotte. 'An Infrastructural Framework to Achieve a European Artificial Intelligence Megaproject', 30 September 2019. https://www.researchgate.net/publication/340574784_An_infrastructural_framework_to_achieve_a_European_artificial_intelligence_megaproject.

[175] Kerry, Cameron F, Joshua P Meltzer, and Andrea Renda. 'AI Cooperation on the Ground: AI Research and Development on a Global Scale'. Brookings Institute & Forum for Cooperation on Artificial Intelligence (FCAI), October 2022. https://www.brookings.edu/wp-content/uploads/2022/11/FCAI-October-2022.pd; Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

[176] Ho and others, 'International Institutions for Advanced AI', p. 14.

large-scale governmental project is infeasible. Moreover, they argue that it is an inappropriate analogy because the Manhattan Project offered a singular goal, whereas AI safety faces a situation where 'ten thousand researchers have ten thousand different ideas on what it means and how to achieve it.'[177]

Model 7: Distribution of benefits and access

7.1 Functions and types: The function of this institutional model is to provide access to the benefits of a technology or a global public good to those States or individuals who do not yet have it due to geographic or economic reasons, among others. Very often, the aim of such an institution is to facilitate *unrestricted access*, or even access schemes targeted to the most needy and deprived. When the information or goods being shared can potentially pose a risk or be misused, yet responsible access is still considered a legitimate, necessary or beneficial goal, institutions under this model tend to create a system for *conditional access*.

7.2 Common examples: Examples of unrestricted benefit-distributor institutions include: international public-private partnerships like Gavi, the Vaccine Alliance, and the Global Fund to Fight AIDS, Tuberculosis and Malaria.[178] Examples of conditional benefit-distributor institutions might include the IAEA's nuclear fuel bank.[179]

7.3 Underexplored examples: Examples that are not yet invoked but that could be promising include the Nagoya Protocol's Access and Benefit Clearing-House (ABS Clearing-House),[180] the UN Climate Technology Centre and Network,[181] and the United Nations Industrial Development Organization (UNIDO), which is tasked with helping build up industrial capacities in developing countries.

7.4 Proposed AI institutions along this model: Stafford and Trager draw an analogy between the NPT and a **potential international regime** to govern transformative AI. The basis for this comparison is that both technologies are dual-use, both present risks even in civilian applications, and there are significant gaps in the access different States have to these technologies. Just like in the case of nuclear energy, in a scenario where there is a clear leader in the race to develop AI and others are lagging, it is mutually beneficial for the actors to enter a technology-sharing bargain. This way, the leader can ensure it will continue to be at the front of the race, while the laggards secure access to the technology. They call this the 'Hopeless Laggard effect'. To enforce this technology-sharing bargain in the sphere of transformative AI, an international

[177] Kaushik, Divyansh, and Matt Korda. 'Panic about Overhyped AI Risk Could Lead to the Wrong Kind of Regulation'. Vox, 3 July 2023. https://www.vox.com/future-perfect/2023/7/3/23779794/artificial-intelligence-regulation-ai-risk-congress-sam-altman-chatgpt-openai.

[178] Ho and others, 'International Institutions for Advanced AI'.

[179] Ibid.

[180] ABSCH. 'Access and Benefit-Sharing Clearing-House'. Accessed 29 August 2023. https://absch.cbd.int/en/.

[181] Climate Technology Centre & Network'. Accessed 29 August 2023. https://www.ctc-n.org/.

institution that conducts similar functions to the IAEA's Global Nuclear Safety and Security Network, which transfers knowledge from countries with mature nuclear energy programmes to those who are just starting to develop one, would have to be created. As an alternative, the authors suggest the leader in AI could prevent the laggards from engaging in a race by sharing the wealth resulting from transformative AI.[182]

The US's National Security Commission on Artificial Intelligence's final report included a proposal for an **International Digital Democracy Initiative** (ISTS) "with allies and partners to align international assistance efforts to develop, promote, and fund the adoption of AI and associated technologies that comports with democratic values and ethical norms around openness, privacy, security, and reliability."[183]

Ho and others envision a model that incorporates the private sector into the benefit-distribution dynamic. A '**Frontier AI Collaborative**' could spread the benefits of cutting-edge AI—including global resilience to the misused or misaligned AI systems—by acquiring or developing AI systems with a pool of resources from Member States and international development programs, or AI laboratories. This form of benefit-sharing could have the additional advantage of incentivizing States to join an international AI governance regime in exchange for access to the benefits distributed by the Collaborative.[184] More broadly, the Elders have recently suggested creating an **institution analogous to the IAEA** to guarantee that AI's benefits are "shared with poorer countries".[185] In forthcoming work, Adan sketches key features for a **Fair and Equitable Benefit Sharing Model**, to "foster inclusive global collaboration in transformative AI development and ensure that the benefits of AI advancements are equitably shared among nations and communities".[186]

7.5 Critiques of this model: One challenge faced by benefit-distributor institutions is how to balance the risk of proliferation with ensuring meaningful benefits and take-up from its technology-promotional and distributive work.[187] For instance, Ho and others

[182] Stafford, Eoghan, and Robert F Trager. 'The IAEA Solution: Knowledge Sharing to Prevent Dangerous Technology Races', 2022, 91. https://www.governance.ai/research-paper/knowledge-sharing-to-prevent-dangerous-technology-races

[183] National Security Commission on Artificial Intelligence. 'Final Report'. Chapter 15.

[184] Ho and others, 'International Institutions for Advanced AI'.

[185] Robinson, Mary. 'The Elders Urge Global Co-Operation to Manage Risks and Share Benefits of AI', 31 May 2023. https://theelders.org/news/elders-urge-global-co-operation-manage-risks-and-share-benefits-ai.

[186] Adan, Sumaya Nur. 'Crucial Features of Fair and Equitable Benefit sharing model for Transformative Artificial Intelligence' (Draft).

[187] To some extent, an effective technology-sharing or access-providing function may be key even for other institutional models that are focused on non-proliferation, insofar as they provide incentives for participation and support different stakeholders (from diplomats to national exports) to come together around a shared mission. See also Roehrlich, Elisabeth. *Inspectors for Peace: A History of the International Atomic Energy Agency*. Johns Hopkins University Press, 2022. https://doi.org/10.1353/book.100164. We thank Harry Law for this observation.

have suggested that proposals such as their Frontier AI Collaborative proposal could face challenges in inadvertently diffusing dangerous dual-use technologies, while simultaneously encountering barriers and obstacles to effectively empowering underserved populations with AI.[188]

More fundamentally, potential challenges or concerns with global benefit- and access-providing institutions—*especially* those that involve some forms of conditional access—will likely see challenges (and critiques) on the basis of how they organize participation. In recent years, several researchers have argued that the global governance of AI is seeing only limited participation by States from the Global South;[189] Veale and others have recently critiqued many initiatives to secure 'AI for Good' or 'responsible AI', arguing that these have fallen into a 'paradox of participation', one involving "the surface-level participation of Global South stakeholders without providing the accompanying resources and structural reforms to allow them to be involved meaningfully".[190] It is likely that similar critiques will be raised against benefit-distributing institutions.

II. Directions for Further Research

In light of the literature review conducted in Part I we can consider a range of additional directions for further research. Without intending to be exhaustive, this section discusses some of those directions briefly, offering some initial thoughts on the existing gaps in current literature, and how each line of research might be helpful to inform key decisions around the international governance of AI—around whether or when to create international institutions; what specific institutional models to prioritize; how to establish these institutions; and how to design them for effectiveness, amongst others.

Direction 1: Effectiveness of institutional models

In the above summary, we have outlined potential institutional models for AI without always making an assessment of their weaknesses or their effectiveness in meeting their stated goals. We believe such further analysis could be critical however in order to filter out models that would be apt to govern the risks from AI, and reduce such risks *de facto* (not just de jure).

[188] Ho and others, 'International Institutions for Advanced AI', p. 12.

[189] Garcia, Eugenio V. 'The Technological Leap of AI and the Global South: Deepening Asymmetries and the Future of International Security'. SSRN Scholarly Paper. Rochester, NY, 10 November 2022. https://doi.org/10.2139/ssrn.4304540.

[190] Veale, Michael, Kira Matus, and Robert Gorwa. 'AI and Global Governance: Modalities, Rationales, Tensions'. Annual Review of Law and Social Science 19, no. 1 (2023): https://doi.org/10.1146/annurev-lawsocsci-020223-040749. Pg. 21. Citing Png, Marie-Therese. 'At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance'. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1434–45. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3531146.3533200.

There is, of course, a live debate on the 'effectiveness' of international law and institutions, with an extensive literature that tries to assess patterns of State compliance with different regimes in international law,[191] as well as more specific patterns affecting the effectiveness of international organizations,[192] or their responsiveness to shifts in the underlying problem.[193]

Such work has highlighted the imperfect track record of many international treaties in meeting their stated purposes,[194] the various ways in which States may aim to evade obligations even while complying with the letter of the law,[195] the ways in which states may aim to use international organizations to promote narrow national interests rather than broader organizational objectives;[196] and the situations under which States aim to exit, shift away from, or replace existing institutions with new replacements.[197] Against such work, other studies have explored the deep normative changes that international norms have historically achieved in topics such as the role of territorial war;[198] studies of transnational and domestic mechanisms by which States are pushed to commit to- and comply with different treaties;[199] or the more nuanced conditions that may induce

[191] See e.g. in the field of arms control, Williamson, Richard. 'Hard Law, Soft Law, and Non-Law in Multilateral Arms Control: Some Compliance Hypotheses'. *Chicago Journal of International Law* 4, no. 1 (1 April 2003). https://chicagounbound.uchicago.edu/cjil/vol4/iss1/7.; Wunderlich, Carmen, Harald Müller, and Una Jakob. 'WMD Compliance and Enforcement in a Changing Global Context'. The United Nations Institute for Disarmament Research, 20 May 2020. https://doi.org/10.37559/WMD/21/WMDCE02.

[192] Coen, David, Julia Kreienkamp, Alexandros Tokhi, and Tom Pegram. 'Making Global Public Policy Work: A Survey of International Organization Effectiveness'. Global Policy 13, no. 5 (2022): 656–68. https://doi.org/10.1111/1758-5899.13125.

[193] Lundgren, Magnus, Jonas Tallberg, Thomas Sommerer, and Theresa Squatrito. 'When Are International Organizations Responsive to Policy Problems?' International Studies Quarterly 67, no. 3 (10 March 2023). https://academic.oup.com/isq/article/67/3/sqad045/7203620.

[194] Hoffman, Steven J., Prativa Baral, Susan Rogers Van Katwyk, Lathika Sritharan, Matthew Hughsam, Harkanwal Randhawa, Gigi Lin, et al. 'International Treaties Have Mostly Failed to Produce Their Intended Effects'. *Proceedings of the National Academy of Sciences* 119, no. 32 (9 August 2022): e2122854119. https://doi.org/10.1073/pnas.2122854119.

[195] Búzás, Zoltán I. 'Evading International Law: How Agents Comply with the Letter of the Law but Violate Its Purpose'. *European Journal of International Relations* 23, no. 4 (1 December 2017): 857–83. https://doi.org/10.1177/1354066116679242.

[196] Lall, Ranjit. 'Beyond Institutional Design: Explaining the Performance of International Organizations'. International Organization 71, no. 2 (ed 2017): 245–80. https://doi.org/10.1017/S0020818317000066.

[197] Eilstrup-Sangiovanni, Mette, and Daniel Verdier. 'To Reform or to Replace? : Institutional Succession in International Organizations'. Working Paper. European University Institute, 2021. https://cadmus.eui.eu/handle/1814/69862.

[198] Hathaway, Oona A., and Scott J. Shapiro. 'International Law and Its Transformation through the Outlawry of War'. International Affairs 95, no. 1 (1 January 2019): 45–62. https://doi.org/10.1093/ia/iiy240.

[199] Hathaway, Oona. 'Between Power and Principle: An Integrated Theory of International Law'. *University of Chicago Law Review* 72, no. 2 (1 March 2005). https://chicagounbound.uchicago.edu/uclrev/vol72/iss2/2.

greater or lesser State compliance with norms or treaties;[200] the effective role that even non-binding norms may play;[201] as well as arguments that a narrow focus on State compliance with international rules understates the broader effects that those obligations may have on the way that States bargain in light of those norms (even when they aim to bend them).[202] Likewise, there is a larger body of foundational work that considers whether traditional international law, based in State consent, might be an adequate tool to secure global public goods such as those around AI, even if States complied with their obligations.[203]

Work to evaluate the (prospective) effectiveness of international institutions on AI could draw on this widespread body of literature to learn lessons from the successes and failures of past regimes, as well as scholarship on the appropriate design of different bodies,[204] measures to improve the decision-making performance of such organizations,[205] to understand when or how any given institutional model might be most appropriately designed for AI.

[200] Kaplow, Jeffrey M. 'State Compliance and the Track Record of International Security Institutions: Evidence from the Nuclear Nonproliferation Regime'. *Journal of Global Security Studies* 7, no. 1 (1 March 2022): ogab027. https://doi.org/10.1093/jogss/ogab027. (on the impact of emerging evidence of increasing noncompliance on the compliance of other States); Neumayer, Eric. 'Do International Human Rights Treaties Improve Respect for Human Rights?' *Journal of Conflict Resolution* 49, no. 6 (1 December 2005): 925–53. https://doi.org/10.1177/0022002705281667.

[201] Shelton, Dinah L, ed. *Commitment and Compliance: The Role of Non-Binding Norms in the International Legal System*. Oxford University Press, 2003.

[202] Howse, Robert, and Ruti Teitel. 'Beyond Compliance: Rethinking Why International Law Really Matters'. *Global Policy* 1, no. 2 (2010): 127–36. https://doi.org/10.1111/j.1758-5899.2010.00035.x.; Meyer, Timothy. 'How Compliance Understates Effectiveness'. *AJIL Unbound* 108 (ed 2014): 93–98. https://doi.org/10.1017/S239877230000194X.

[203] Aaken, Anne van. 'Is International Law Conducive To Preventing Looming Disasters?' *Global Policy* 7, no. S1 (2016): 81–96. https://doi.org/10.1111/1758-5899.12303.; See generally Trachtman, Joel P. The Future of International Law: Global Government. ASIL Studies in International Legal Theory. Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9781139565585. See also: Pauwelyn, J., R. A. Wessel, and J. Wouters. 'When Structures Become Shackles: Stagnation and Dynamics in International Lawmaking'. European Journal of International Law 25, no. 3 (1 August 2014): 733–63. https://doi.org/10.1093/ejil/chu051. For a critique of this trend, see Krisch, Nico. 'The Decay of Consent: International Law in an Age of Global Public Goods'. American Journal of International Law 108, no. 1 (January 2014): 1–40. https://doi.org/10.5305/amerjintelaw.108.1.0001.

[204] Ulfstein, Geir. 'Reflections on Institutional Design – Especially Treaty Bodies'. Research Handbook on the Law of International Organizations, 29 July 2011. https://www.elgaronline.com/view/edcoll/9781847201355/9781847201355.00024.xml.

[205] Sommerer, Thomas, Theresa Squatrito, Jonas Tallberg, and Magnus Lundgren. 'Decision-Making in International Organizations: Institutional Design and Performance'. The Review of International Organizations 17, no. 4 (1 October 2022): 815–45. https://doi.org/10.1007/s11558-021-09445-x.

Direction 2: Multilateral AI treaties without institutions

While our review has focused on international AI governance proposals that would involve the establishment of some forms of international institutions, there are of course other models of international cooperation. Indeed, some types of treaties, do not automatically establish distinct international organizations,[206] and primarily function by setting shared patterns of expectations and reciprocal behavior amongst states, in order (ideally) to become self-enforcing. As discussed, our literature review omits discussing this type of regime. However, analyzing them in combination with the models we have outlined could be useful to determine international governance alternatives for AI, including whether or when state initiatives to establish such multilateral normative regimes that lack treaty bodies would likely be effective, or might likely fall short.

Such an analysis could draw from a rich vein of existing proposals for new international treaties on AI. There have of course been proposals for new treaties for autonomous weapons.[207] There are also proposals for international conventions to mitigate extreme risks from technology. Some of these, such as Wilson's '**Emerging Technologies Treaty**',[208] or Verdirame's **Treaty on Risks to the Future of Humanity**,[209] would address many types of potential existential risks simultaneously, including potential risks from AI.

Other treaty proposals are focused more specifically on regulating AI risk in particular. Dewey discusses a potential '**AI development convention**' that would set down 'a ban or set of strict safety rules for certain kinds of AI development.'[210] Yet others address different types of risks from AI, such as Metzinger's proposal for a **global moratorium on artificial suffering**.[211] Carayannis and Draper have discussed a '**Universal Global Peace Treaty**' **(UGPT)**, which would commit States 'not to declare or engage in interstate war, especially via existential warfare, i.e., nuclear, biological, chemical, or

[206] For discussion of the conditions under which States may choose to establish treaty bodies rather than full-fledged intergovernmental organizations, as well as design options for treaty bodies, see also: Ulfstein, Geir. 'Treaty Bodies and Regimes'. Oxford Guide to Treaties, Duncan B. Hollis, ed., 2012. https://doi.org/10.2139/ssrn.2144650.

[207] Docherty, Bonnie. '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.

[208] Wilson, Grant. 'Minimizing Global Catastrophic and Existential Risks from Emerging Technologies through International Law'. Va. Envtl. LJ 31 (2013): 307. http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/velj31&section=12

[209] Verdirame, Guglielmo. 'For China, a Legal Reckoning Is Coming'. UnHerd, 20 April 2020. https://unherd.com/2020/04/for-china-a-legal-reckoning-is-coming/.

[210] Dewey, Daniel. 'Long-Term Strategies for Ending Existential Risk from Fast Takeoff'. In Risks of Artificial Intelligence. Chapman and Hall/CRC, 2015. https://www.taylorfrancis.com/chapters/edit/10.1201/b19187-14/long-term-strategies-ending-existential-risk-fast-takeoff-daniel-dewey. Pg. 7-8.

[211] Metzinger, Thomas. 'Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology'. Journal of Artificial Intelligence and Consciousness, 19 February 2021, 1–24. https://doi.org/10.1142/S270507852150003X.

cyber war, including AI- or ASI enhanced war'. They would see this treaty supported by a separate **Cyberweapons and AI Convention**, which would commit as main article that 'each State Party to this Convention undertakes never in any circumstances to develop, produce, stockpile or otherwise acquire or retain: (1) cyberweapons, including AI cyberweapons; and (2) AGI or artificial superintelligence weapons.'[212]

Notwithstanding these proposals, there are significant gaps in the scholarship surrounding the design of an international treaty for AI regulation. Some issues that we believe should be explored include, but are not limited to: the effects of reciprocity on the behavior of States Parties; the relationship between the specificity of a treaty and its pervasiveness; the success and adaptability of the framework convention model (a broad treaty and protocols which specify the initial treaty's obligations) in accomplishing their goals; adjudicatory options for conflicts between States Parties.

Direction 3: Additional institutional models not covered in detail in this review

There are many other institutional models that this literature review does not address, as they are (currently) rarely proposed in the specific context of international AI governance. These include, but are not limited to:[213]

- Various **International non-governmental organizations (NGOs)**, e.g. the World Wide Fund for Nature (WWF), Amnesty International (AI);
- **Political and economic unions:** e.g. Association of Southeast Asian Nations (ASEAN),
- **Military alliances** that establish security guarantees and/or political, economic, and defense cooperation, e.g. North Atlantic Treaty Alliance (NATO); the Shanghai Cooperation Organisation (SCO), or the Collective Security Treaty Organization (CSTO).
- **International courts and tribunals**, e.g. the International Criminal Court (ICC), various regional courts of human rights (African Court on Human and Peoples' Rights, the European Court of Human Rights, the Inter-American Court of Human Rights);

[212] Carayannis, Elias G., and John Draper. 'Optimising Peace through a Universal Global Peace Treaty to Constrain the Risk of War from a Militarised Artificial Superintelligence'. AI & SOCIETY, 11 January 2022. https://doi.org/10.1007/s00146-021-01382-y.

[213] For another review of functions, see also Sepasspour, Rumtin. 'A Reality Check and a Way Forward for the Global Governance of Artificial Intelligence'. *Bulletin of the Atomic Scientists*, 10 September 2023. https://www.tandfonline.com/doi/abs/10.1080/00963402.2023.2245249. ('At a high level, these functions can be principles and guidelines; goals, metrics and targets; data collection and reporting; research and development; forecasting and horizon-scanning; forums and convening; norms and standards; rules; laws and legal conventions; funding; capacity-building; direct assistance, such as aid relief; certification; monitoring, verification and auditing; revenue collection; dispute settlement and arbitration; adjudication; sanctions; and enforcement.')

- **Inter-State arbitral & dispute settlement bodies**, e.g. the International Court of Justice (ICJ); the WTO Appellate Body which hears disputes by WTO Members; the International Tribunal for the Law of the Sea (ITLOS), which is one of the dispute resolution mechanisms for the UN Convention on the Law of the Sea (UNCLOS); the Permanent Court of Arbitration (PCA), which resolves disputes arising out of international agreements between Member States, international organizations, or private parties; the European Nuclear Energy Tribunal, which oversees nuclear energy disputes within the OECD;

- **Cartels** aimed at articulating, aggregating and securing the (economic) interests of their members, e.g. Organization of the Petroleum Exporting Countries (OPEC), whose members cooperate to reduce market competition but whose operations may be protected by the doctrine of State immunity under international law;

- **Policy Implementation and/or direct service delivery organizations**, e.g. the the United Nations Development Programme (UNDP), or the World Bank;

- **Data gathering & dissemination organizations**, e.g. the World Meteorological Organization's (WMO) climate data monitoring; the Food and Agriculture Organization (FAO) gathering of statistics on global food production;

- **Post-disaster response & relief organizations**, e.g. The World Food Programme (WFP) or the International Committee of the Red Cross (ICRC);

- **Capacity-building & training organizations**, e.g. governmental capacity-building trainings offered by the United Nations Institute for Training and Research (UNITAR); fiscal management training programs offered by the International Monetary Fund (IMF); border control trainings provided by the International Organization for Migration (IOM);

- **Norm promotion organizations**: e.g. the UNESCO World Heritage site program; UNHCR advocacy for refugee rights;

- **Awareness-raising organizations**, e.g. the Joint United Nations Programme on HIV/AIDS (UNAIDS), which amongst others organizes World AIDS Day; and

- **'Meta'-organizations** which aim to support or enhance the activities of other existing international organizations in general; e.g. the United Nations Office for Project Services (UNOPS).

Accordingly, future lines of research could focus on exploring what such models could look like for AI, and what they might contribute to international AI governance.

Direction 4: Compatibility of institutional functions

There are multiple instances of compatibility between the functions of institutions proposed by the literature explored in this review. Getting a better sense of those areas of compatibility could be advantageous when designing an international institution for

AI that borrows the best features from each model rather than copying a single model. Further research could explore hybrid institutions that combine functions from several models. Some potential combinations include, but are not limited to:

- **Comprehensive scientific consortia**, which could combine elements from scientific consensus-building institutions, international joint scientific programs, and (scientific) benefit-distributing institutions;
- **Full-spectrum consensus-building fora,** which could combine elements from scientific consensus-building with political consensus-building institutions, and potentially with stabilization and emergency-response institutions;
- **Integrated regulator institutions**, which could combine elements from regulatory and policy coordinator institutions with monitoring and verification institutions; and
- **Centralized control institutions**, which could combine elements from nonproliferation, export control institutions, with access-controlling institutions, and potentially with monitoring and verification institutions.

Direction 5: Potential fora for an international AI organization

This review omits establishing patterns among different proposals on their preferred fora to negotiate or host an international AI organization. While we do not expect there to be much commentary on this, it might be a useful additional element to take into consideration when designing an international AI institution. For example, some fora that have been proposed are:

- The **United Nations** could establish a UN specialized agency through State negotiations or initiative at the UN General Assembly.[214] For instance, as seen above, Kemp and others call for a UN AI Research Organization (UNAIRO).[215]
- **Regional organizations**, such as the European Union, the Organization of American States, the African Union, or ASEAN, could pioneer regional regulatory regimes that exert indirect extraterritorial effects on global AI governance. The European Union in particular has proven to be effective at indirectly regulating industries at a global level through the so-called Brussels

[214] See Robinson, Mary. 'The Elders Urge Global Co-Operation to Manage Risks and Share Benefits of AI', 31 May 2023. https://theelders.org/news/elders-urge-global-co-operation-manage-risks-and-share-benefits-ai.

[215] Kemp, Luke, Peter Cihon, Matthijs Michiel Maas, Haydn Belfield, Zoe Cremer, Jade Leung, and Seán Ó hÉigeartaigh. 'UN High-Level Panel on Digital Cooperation: A Proposal for International AI Governance'. Centre for the Study of Existential Risk and Leverhulme Centre for the Future of Intelligence, 26 February 2019. https://www.cser.ac.uk/news/advice-un-high-level-panel-digital-cooperation/.

Effect.[216] Siegmann and Anderljung suggest the EU AI Act could have a similar effect on the global AI industry.[217]

- Similarly, **minilateral club organizations** like the G7, BRICS, the G20 or the OECD could play a similar role, bringing together like-minded countries to negotiate an international governance framework for AI that other States can then join.[218]

- **Public-private partnerships or coalitions** between state and non-state actors, such as the Lysøen Initiative on human security[219] or the Christchurch Call, an initiative (led by France and Aotearoa New Zealand) on eliminating online terrorist and violent extremist content,[220] which can organize a coalition of like-minded States and actors to pursue the negotiation of new treaties, where necessary outside of UN fora.

- **Gradual formalization of initial informal institutions**: in some cases, organizations that are initially established in an informal configuration could lay the foundation for formal frameworks for cooperation, as happened with the General Agreement on Tariffs and Trade's (GATT) gradual transformation into the WTO, and which Erdélyi and Goldsmith have suggested as one route that could be taken by an International Artificial Intelligence Organization.[221]

This does not exhaust the available or feasible avenues, however. In many cases, significant additional work will have to be undertaken to evaluate these pathways in detail.

[216] Bradford, Anu. *The Brussels Effect: How the European Union Rules the World*. Oxford, New York: Oxford University Press, 2020.

[217] Siegmann, Charlotte, and Markus Anderljung. 'The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market'. Centre for the Governance of AI, August 2022. https://www.governance.ai/research-paper/brussels-effect-ai.

[218] Morin, Jean‐Frédéric, Hugo Dobson, Claire Peacock, Miriam Prys‐Hansen, Abdoulaye Anne, Louis Bélanger, Peter Dietsch, et al. 'How Informality Can Address Emerging Issues: Making the Most of the G7'. Global Policy 10, no. 2 (May 2019): 267–73. https://doi.org/10.1111/1758-5899.12668.

[219] Maas, 'Artificial Intelligence Governance Under Change', pg 308.; Basu and Sherman, 'Two New Democratic Coalitions on 5G and AI Technologies'.

[220] The Christchurch Call. 'Home'. Christchurch Call. Accessed 18 September 2023. https://www.christchurchcall.com/. See also: Veale, Kevin. 'Conclusion: The Christchurch Call to Action Summit and What Follows'. In Gaming the Dynamics of Online Harassment, edited by Kevin Veale, 147–63. Cham: Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-60410-3_7.

[221] Erdélyi, Olivia J., and Judy Goldsmith. 'Regulating Artificial Intelligence: Proposal for a Global Solution'. Pg. 14.

Conclusion

This literature review analyzed seven models for the international governance of AI, discussing common examples of those models, underexplored examples, specific proposals of their application to AI in existing scholarship, and critiques. We found that, while the literature covers a wide range of options for the international governance of AI, most of the time proposals are vague and do not develop the specific attributes that an international institution would need to have to garner the benefits and curb the risks associated with AI. Thus, we proposed a series of pathways for further research that we expect should contribute to the design of such an international institution.

Legal Priorities
Project

www.legalpriorities.org