Global AI governance: barriers and pathways forward

HUW ROBERTS, EMMIE HINE, MARIAROSARIA TADDEO AND LUCIANO FLORIDI*

Late 2022 and early 2023 saw the commercialization of powerful new artificial intelligence (AI) technologies such as OpenAI's ChatGPT. These systems have numerous benefits, including improving business efficiency and enhancing consumer experiences, but also pose significant risks. They threaten national security by democratizing capabilities that could be used by malicious actors; facilitate unequal economic outcomes by concentrating market power in the hands of a few companies and countries, while displacing jobs in others; and produce societally undesirable conditions through extractive data practices, reinforcing biased narratives and environmentally harmful compute requirements.[1]

These risks transcend national borders and have reinvigorated calls for stronger global AI governance, understood here as the process through which diverse interests that transcend borders are accommodated, without a single sovereign authority, so that cooperative action may be taken in maximizing the benefits and mitigating the risks of AI.[2] The United Nations Secretary-General António Guterres, British Prime Minister Rishi Sunak and OpenAI CEO Sam Altman have all argued for the creation of a new international AI body modelled on existing institutions like the Intergovernmental Panel on Climate Change (IPCC) and the International Atomic Energy Agency (IAEA). A new-found emphasis on global AI governance is promising, but this type of ambitious governance proposal is generally misaligned with current geopolitical and institutional realities, raising questions over desirability and feasibility.

This policy paper is a response to the growing calls for ambitious new international institutions for AI.[3] It maps the geopolitical and institutional barriers to stronger global AI governance and considers potential pathways forward in

[*] This article is part of *International Affairs*' policy papers series—a forum for bringing new insights into policy debates, for rapidly publishing new empirical results and for developing potential solutions to international problems. Huw Roberts' research was supported by a studentship from the Grand Union Doctoral Training Partnership.

[1] Lewis Ho et al., 'International institutions for advanced AI', arXiv, 11 July 2023, https://doi.org/10.48550/arXiv.2307.04699.

[2] Lawrence S. Finkelstein, 'What is global governance?', *Global Governance* 1: 3, 1995, pp. 367–72; Thomas G. Weiss, 'Governance, good governance and global governance: conceptual and actual challenges', *Third World Quarterly* 21: 5, 2000, pp. 795–814, https://doi.org/10.1080/713701075.

[3] Matthijs M. Maas and José Jaime Villalobos, 'International AI institutions: a literature review of models, examples, and proposals', Legal Priorities Project, 22 Sept 2023, https://doi.org/10.2139/ssrn.4579773.

International Affairs **100**: 3 (2024) 1275–1286; DOI: 10.1093/ia/iiae073

light of these constraints. We argue that a promising foundation of international regimes focused on AI governance is emerging, but the centrality of AI to interstate competition, dysfunctional international institutions and disagreement over policy priorities problematizes substantive cooperation. We propose strengthening the existing weak 'regime complex' of international institutions as the most desirable and realistic path forward for global AI governance. Strengthening coordination between, and the capacities of, existing institutions supports mutually reinforcing policy change, which, if enacted properly, can lead to catalytic change across the various policy areas where AI has an impact. It also facilitates the flexible governance needed for rapidly evolving technologies.

To make this argument, we outline key global AI governance processes in the next section. In the third section, we analyse how first- and second-order cooperation problems in international relations apply to AI. In the fourth section we assess potential routes for advancing global AI governance, and we conclude by providing recommendations on how to strengthen the weak AI regime complex.

The landscape of global AI governance initiatives

States and international institutions have been active in developing international AI governance initiatives. For instance, discussions have taken place in the UN since 2014 about governing lethal autonomous weapons systems (LAWS) under the Convention on Certain Conventional Weapons.[4] In 2019 OECD member countries adopted a set of AI ethics principles, with G20 leaders subsequently committing to principles drawn from the OECD set.[5] In November 2021, all 193 of UNESCO's member states adopted a Recommendation on the Ethics of Artificial Intelligence, designed to guide signatories in developing appropriate legal frameworks.[6] Then, in 2023, the G7 initiated the Hiroshima AI Process to enhance cooperation in AI governance,[7] while the BRICS countries (Brazil, Russia, India, China and South Africa) agreed to form an 'AI study group'.[8] Finally, the Council of Europe (CoE) has been developing a legally binding international convention on AI and human rights, with a draft text published in December 2023.[9]

Efforts have also been made to establish new international AI bodies. The Global Partnership on AI (GPAI), launched in 2020 by 15 founding countries to support the ethical adoption of AI, is one example.[10] The Trade and Technology

[4] United Nations Office for Disarmament Affairs, 'Timeline of LAWS in the CCW', https://disarmament.unoda.org/timeline-of-laws-in-the-ccw/. (Unless otherwise noted at point of citation, all URLs cited in this article were accessible on 20 February 2024.)

[5] OECD.AI, 'About OECD.AI', https://oecd.ai/en/about/background.

[6] United National Educational, Scientific and Cultural Organization (UNESCO), *Recommendation on the Ethics of Artificial Intelligence* (Paris: UNESCO, 2022), https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[7] European Commission, 'G7 leaders' statement on the Hiroshima AI Process', 30 Oct. 2023, https://digital-strategy.ec.europa.eu/en/library/g7-leaders-statement-hiroshima-ai-process.

[8] Ethan Wang and Liz Lee, 'China's Xi calls for accelerated BRICS expansion', Reuters, 23 Aug. 2023, https://www.reuters.com/world/chinas-xi-calls-accelerated-brics-expansion-2023-08-23/.

[9] Council of Europe Committee on AI, *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law,* 18 Dec.2023, https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043.

[10] The Global Partnership on AI, 'About GPAI', https://gpai.ai/about/.

Council, established in 2021 to coordinate European Union and United States government activities in trade and technology, including AI, is another.[11] In 2023, the UN Secretary-General's envoy on technology announced the creation of a High-Level Advisory Body on AI, tasked with advancing recommendations for international AI governance,[12] and the United Kingdom unilaterally established an AI Safety Institute designed to advance global knowledge on advanced AI.[13]

Voluntary initiatives have struggled to address the myriad harms from AI. High-level principles, such as those agreed by the G20, are vague and accommodate different ideological positions. Take AI fairness, a principle supported by all G20 members, as applied to facial recognition technology. The implementation of this principle in the EU involves banning these technologies, while in China ethnic-recognition technologies are permissible in the name of social stability.[14] Efforts to build on these thin normative foundations have faced challenges. For instance, since UNESCO's AI Recommendation was adopted in November 2021, fewer than a quarter of signatories have worked with the body to implement proposed policy tools.[15] Signatories have faced no repercussions for non-implementation because the agreement is non-binding. Even in new AI-specific institutions with a narrower membership, like the GPAI, little progress has been made in reaching agreement on meaningful governance initiatives.

The CoE's draft convention on AI holds more promise, as negotiations are at an advanced stage, and ratifying states would be expected to translate the convention into domestic law. However, the ratification of CoE conventions has historically been slow, which is problematic given the rapid pace of AI development. States not involved in drafting CoE conventions have also previously refused to ratify because of a perceived lack of legitimacy, instead pushing for a more representative UN-led process.[16]

Private stakeholders have also developed AI governance mechanisms. International standards bodies, including the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), have published product and process standards for AI.[17] These standards are voluntary, yet some may be mandated as *de jure* legal requirements in legislation or become industry best practice and require *de facto* compliance.[18] Industry has also created

[11] European Commission, 'EU–US Trade and Technology Council', https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en.

[12] United Nations, 'About the Advisory Body on Artificial Intelligence (AI)', https://www.un.org/en/ai-advisory-body/about.

[13] Department for Science, Innovation and Technology, 'Introducing the AI Safety Institute', 17 Jan. 2024, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

[14] Huw Roberts et al., 'Governing artificial intelligence in China and the European Union: comparing aims and promoting ethical outcomes', *The Information Society* 39: 2, 2023, pp. 79–97, https://doi.org/10.1080/01972243.2022.2124565.

[15] UNESCO, 'Artificial intelligence: UNESCO calls on all governments to implement global ethical framework without delay', 30 March 2023, https://www.unesco.org/en/articles/artificial-intelligence-unesco-calls-all-governments-implement-global-ethical-framework-without.

[16] Jonathan Clough, 'The Council of Europe Convention on Cybercrime: defining 'crime' in a digital world', *Criminal Law Forum* 23: 4, 2012, pp. 363–91, https://doi.org/10.1007/s10609-012-9183-3.

[17] The AI Standards Hub, 'Standards database', https://aistandardshub.org/ai-standards-search/.

[18] Tim Büthe and Walter Mattli, *The new global rulers: the privatization of regulation in the world economy* (Princeton: Princeton University Press, 2011).

new governance institutions that set international rules. The Partnership on AI (PAI) was established in 2016 by 'big tech' companies, civil society organizations and academic stakeholders to develop guidance and inform public policy.[19] The Frontier Model Forum was founded by four big tech companies in 2023 specifically to establish good governance mechanisms for advanced systems.[20]

International standards bodies have made progress in developing governance initiatives, yet their primary focus is producing standards that support regulatory consistency which allow organizations to scale.[21] There are signs of these institutions adopting a wider remit for AI, with the EU's draft AI Act relying on standards bodies to explicate value-laden governance issues.[22] However, such provisions are generally not in companies' commercial interests, leading to a 'principal moral hazard' because companies have least to gain from developing stringent protections.[23] This incentive problem also undermines privately led governance initiatives like the PAI. Indeed, the digital rights organization Access Now resigned from the PAI in 2020 after concluding that the organization had failed to influence the attitude of member companies.[24]

This global AI governance landscape can be conceptualized as a weak 'regime complex' because it has a 'polycentric' structure with some linkages between institutions—such as the G7's Hiroshima Process drawing on work from the OECD and GPAI—but work is generally siloed. For some aspects of AI policy, this is relatively unproblematic. AI is not a single-policy problem, but rather a set of loosely connected problems that arise from introducing autonomous agents to tasks. Because of this, it is reasonable to expect little coordination between, for example, the UN's efforts to develop rules for LAWS and technical standards bodies' work to develop AI risk management processes for businesses. In other areas where coordination would be beneficial, such as in the development of authoritative AI principles, siloed efforts have led to detrimental fragmentation.[25]

Barriers to strong global AI governance

The weak AI regime complex has left a governance deficit due to the inadequacy of existing initiatives, gaps in the landscape and difficulties reaching agreement

[19] Partnership on AI, 'About Us', https://partnershiponai.org/about/.

[20] Google, 'Frontier Model Forum: a new partnership to promote responsible AI', 26 July, 2023, https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/

[21] Peter Cihon, *Standards for AI governance: international standards to enable global coordination in AI research & development* (Oxford: Future of Humanity Institute, 2019), https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_-FHI-Technical-Report.pdf.

[22] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the draft EU Artificial Intelligence Act—analysing the good, the bad, and the unclear elements of the proposed approach', *Computer Law Review International* 22: 4, 2021, pp. 97–112, https://doi.org/10.9785/cri-2021-220402.

[23] Gary J. Miller and Andrew B. Whitford, 'The principal's moral hazard: constraints on the use of incentives in hierarchy', *Journal of Public Administration Research and Theory* 17: 2, 2007, pp. 213–33, https://doi.org/10.1093/jopart/mul004.

[24] Michael Veale, Kira Matus and Robert Gorwa, 'AI and global governance: modalities, rationales, tensions', *Annual Review of Law and Social Science* 19: 1, 2023, pp. 255–75, https://doi.org/10.1146/annurev-lawsocsci-020223-040749.

[25] Anna Jobin, Marcello Ienca and Effy Vayena, 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, vol. 1, 2019, pp. 389–99, https://doi.org/10.1038/s42256-019-0088-2.

over more suitable mechanisms. As the international AI governance landscape matures, the characteristics of AI mean that first- and second-order cooperation problems[26] will pose significant—though not insurmountable—challenges to developing effective global governance mechanisms for these technologies. First-order cooperation problems are geopolitical challenges that stem from the condition of international anarchy, understood here as 'a lack of common government in world politics, not [as] a denial that an international society—albeit a fragmented one—exists'.[27] In the absence of a sovereign authority, states face uncertainty over the enforcement of agreements and other states' intentions. The degree to which cooperative action takes place under these conditions differs by policy area and is influenced by factors like states' threat perceptions, levels of trust and alignment of interests.

AI is particularly subject to first-order cooperation problems because states perceive it as a source of competitive advantage. Because AI is a dual-use technology, technical breakthroughs can provide economic and security benefits. This is an explicit policy aspiration for China, which seeks to use dual-use technologies to promote military–civil fusion through technology transfer between the sectors.[28] Fears of China gaining advantages from AI have led the US to situate itself as the leader of 'ideologically aligned' countries in opposition to China, including for security purposes.[29]

The perceived centrality of AI for competition has led states to enact policies to strengthen their international position. The US introduced export controls for semiconductors to hinder China's AI development, while also promoting domestic semiconductor production. China has used industrial policies, like Made in China 2025, to promote competitiveness and has become an international norm-maker by exporting technology standards through its Belt and Road Initiative. The EU has been pursuing a policy of 'digital sovereignty' to strengthen domestic hardware and software capacities and lessen reliance on foreign technologies. While arguably incidental, the 'Brussels effect' from EU regulatory efforts gives it influence in technology competition by shaping the rules that companies follow internationally.[30]

Policies supporting national competitiveness are not necessarily detrimental for cooperation. Yet, the perception of AI's centrality for competitive advantages and narratives of an 'arms race' have amplified first-order cooperation problems, undermining mutual trust. This is particularly true for some AI technologies—like cutting-edge 'foundation models'[31]—and between specific countries, notably

[26] Thomas Hale, David Held et al., *Beyond gridlock* (Hoboken, NJ: Wiley, 2017).

[27] Robert Axelrod and Robert O. Keohane, 'Achieving cooperation under anarchy: strategies and institutions', *World Politics* 38: 1, 1985, pp. 226–54, https://doi.org/10.2307/2010357.

[28] Huw Roberts et al., 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', *AI & Society* 36: 1, 2021, pp. 59–77, https://doi.org/10.1007/s00146-020-00992-2.

[29] James Johnson, 'AI-security dilemma: insecurity, mistrust, and misperception under the nuclear shadow', in James Johnson, *AI and the bomb: nuclear strategy and risk in the digital age* (Oxford: Oxford University Press, 2023).

[30] Anu Bradford, *Digital empires: the global battle to regulate technology* (New York: Oxford University Press, 2023).

[31] Foundation models are AI systems trained on broad data which allows them to be used for a variety of tasks. Their versatility and advanced capabilities facilitate competitive advantages.

China and the US. For instance, a Chinese delegate stated at the first UN Security Council meeting focused on AI that US export controls undermined prospects for international cooperation on global AI governance.[32] Some industry figures have leveraged this 'arms race' framing to simultaneously push for more funding and less domestic regulation,[33] heightening safety risks and exacerbating problems of international trust.

First-order cooperation problems can be mitigated by international institutions that provide a framework for cooperation and facilitate communication. However, second-order cooperation problems stemming from institutional dysfunction compound difficulties in establishing effective global AI governance mechanisms. A proliferation of international institutions following the Second World War, combined with breakthroughs in transportation and information technology, deepened integration among states and transformed many domestic policy areas into international ones. However, this success has complicated contemporary multilateral cooperation. Decolonization, facilitated by institutionalization, has resulted in more participants in global governance. Institutional inertia has prevented adaptation to this new reality, while increasing globalized connectivity requires institutions to address more complex problems. Although new international institutions have emerged to address new policy problems, this has arguably exacerbated institutional fragmentation and mandate overlap, limiting the effectiveness of regimes.[34]

The capacity to develop and regulate AI is currently highly concentrated,[35] indicating that multilateral agreement between China, the EU and the US may be sufficient for mitigating key risks. Legitimacy questions aside, there is theoretically significant scope for agreement when compared to more multipolar policy areas, like health. However, there is little consensus regarding necessary policy responses. The EU foregrounds new regulations, the US is taking a more *laissez-faire* approach, and China relies on a hybrid approach that utilizes industry self-discipline and targeted secondary legislation. At an international level, this has led to disagreements in bodies such as the G7 over what types of international governance instruments should be developed, even among those in what is sometimes termed an 'ideological coalition'.[36] This has restricted agreement to high-level interventions which have had limited success.

The complexity of AI as a policy area complicates the reaching of international agreement. There is little consensus among stakeholders as to which problems should be prioritized. For instance, there is disagreement between 'long-termist' scholars, who focus on the potential existential threats posed by AI, and those

[32] United Nations, 'Artificial intelligence: opportunities and risks for international peace and security—Security Council, 9381st meeting', 18 July 2023, https://media.un.org/en/asset/k1j/k1ji81po8p.

[33] Eric Schmidt, 'I used to run Google. Silicon Valley could lose to China', *New York Times*, 27 Feb. 2020, https://www.nytimes.com/2020/02/27/opinion/eric-schmidt-ai-china.html.

[34] Hale, Held et al., *Beyond gridlock*.

[35] Steven Weber, 'Data, development, and growth', *Business and Politics* 19: 3, 2017, pp. 397–423, https://doi.org/10.1017/bap.2017.3.

[36] Paul Samson, 'On advancing global AI governance', Centre for International Governance Innovation, 1 May 2023, https://www.cigionline.org/articles/on-advancing-global-ai-governance.

more concerned with the harms already materializing, such as bias.[37] This divergence is manifesting at the state level, with the UK placing greater emphasis on long-term safety risks than the EU, which has focused on extant harms.

The fragmented international AI landscape further problematizes cooperation problems as it allows countries and companies to follow different policies and standards for AI.[38] For instance, China is not a member of the GPAI, nor are any Chinese companies involved in the PAI, indicating that two separate AI governance ecosystems are forming. Even where both China and western states participate in the same international organizations, mandate overlap weakens the authoritativeness of any one institution. For instance, the ISO and IEC, as well as the Institute of Electrical and Electronics Engineers (IEEE) and the International Telecommunication Union (ITU), have all been active in AI standards-making, enabling companies to 'forum shop'.

Pathways forward

Addressing the global AI governance deficit requires moving from a weak regime complex to the strongest governance system possible under current geopolitical and institutional conditions. We consider two pathways forward: first, developing new centralized global AI institution(s) and second, strengthening coordination between, and capacities of, existing institutions. To assess these options, we focus on the political legitimacy of regimes, understood in terms of acceptability of political power, as this is the essential condition for effective governance. We use two common sources of political legitimacy as criteria for assessing these options. These are:

(1) A regime's ability to enable states and other actors to coordinate their behaviour in mutually beneficial ways;[39] and
(2) The presence of democratic procedures for decision-making.[40]

In assessing the proposed pathways forward, we focus less on idealized institutional solutions and more on whether the pathway can realize benefits considering cooperation problems. Accordingly, we assess each regime type for AI in the abstract, as well under current geopolitical and institutional conditions.

Several proposals for a new international AI body have been made, often based on existing institutions.[41] The most centralized option is to emulate nuclear governance, which relies on the IAEA to set standards, undertake compliance monitoring and control access to materials. Establishing an 'IAEA for AI' with

[37] 'Stop talking about tomorrow's AI doomsday when AI poses risks today', editorial, *Nature* 618: 7967, 2023, pp. 885–6, https://doi.org/10.1038/d41586-023-02094-7.

[38] Peter Cihon, Matthijs M. Maas and Luke Kemp, 'Fragmentation and the future: investigating architectures for international AI governance', *Global Policy* 11: 5, 2020, pp. 545–6, https://doi.org/10.1111/1758-5899.12890.

[39] Allen Buchanan and Robert O. Keohane, 'The legitimacy of global governance institutions', *Ethics & International Affairs* 20: 4, 2006, pp. 405–37, https://doi.org/10.1111/j.1747-7093.2006.00043.x.

[40] Eva Erman and Markus Furendal, 'Artificial intelligence and the political legitimacy of global governance', *Political Studies*, publ. online 3 Oct. 2022, https://doi.org/10.1177/00323217221126665.

[41] Maas and Villalobos, 'International AI institutions'.

similar powers is unlikely to be an effective way of coordinating state action. Although it would mitigate current institutional friction, centralized regime mandates are often brittle, with an AI institution being at particular risk due to the pace of development.[42] More importantly, nuclear and AI are not similar policy problems. AI policy is loosely defined, with disagreement over field boundaries and what constitutes harm. It is decentralized, meaning it does not face the same physical bottlenecks in materials as nuclear. It also has a cross-cutting upstream and downstream impact, as well as across sectors. A centralized regime for AI would require an unprecedented set of powers to address the range of governance issues associated with AI, including access to private sector developments, suggesting it is unviable in practice.

Other proposals for new institutions target specific AI governance issues, including an 'IPCC for AI' to support scientific consensus and a 'CERN[43] for AI' to undertake advanced safety research. A semi-centralized regime organized around a handful of new AI-specific institutions is a more realistic solution that would mitigate some rigidity problems. However, establishing new institutions risks further fragmenting the governance landscape, thus diluting authority. Establishing this type of regime would also face viability challenges as the international bodies offered as exemplars for AI governance were established under very different geopolitical conditions. The IAEA was created in 1957 during a period of proliferation in international institutions and only gained substantive powers after the Non-Proliferation Treaty was concluded in 1968.[44] Consensus on the existential risk posed by nuclear weapons aided these efforts, something which is not present for AI. The IPCC emerged from decades of expert international cooperation that provided the foundation for the body's work, which is also absent for AI.[45] Contemporary first- and second-order cooperation problems complicate this type of formal institution creation.

The second option—strengthening the existing weak AI regime complex—has a different set of benefits and drawbacks in respect to coordinating behaviour. A regime complex model allows for cooperation in different forums, even while geopolitical or institutional conditions stall progress in others. This facilitates incremental progress and trust-building from myriad state and non-state actors that produce mutually reinforcing change over time.[46] It also allows for adaptability in line with technological change and the inclusion of a diversity of governance stakeholders,[47] including big tech, which is necessary given the technical and often contextual nature of AI. There are drawbacks to a regime complex model, notably related to actors shirking responsibility or reneging on promises, as has been seen with government and private sector climate pledges. Nonethe-

[42] Cihon, Maas and Kemp, 'Fragmentation and the future'.

[43] CERN is the European Organization for Nuclear Research.

[44] Michael Clarke, 'Weapons of mass destruction: incremental steps', in Hale, Held et al., *Beyond gridlock*.

[45] Kari De Pryck and Mike Hulme, eds, *A critical assessment of the Intergovernmental Panel on Climate Change* (Cambridge, UK: Cambridge University Press, 2022).

[46] Robert O. Keohane and David G. Victor, 'The regime complex for climate change', *Perspectives on Politics* 9: 1, 2011, pp. 7–23, https://doi.org/10.1017/S1537592710004068.

[47] Cihon, Maas and Kemp, 'Fragmentation and the future'.

less, research has highlighted that the boldest climate pledges have been the most credible,[48] and there has been no shortage of actors willing to act as norm entrepreneurs for AI, indicating a willingness to catalyse change.

The benefits of a regime complex model for AI are currently undermined by a lack of institutional coordination and authoritativeness, which has left the governance landscape fragmented and contradictory. A stronger regime complex would involve a high degree of coordination and coherence between actors, with complementary initiatives supporting a comprehensive approach to governing AI. Moving from a weak to strong regime complex could involve aligning targets, improving information-sharing, developing institutional partnerships and creating conflict resolution mechanisms.[49] It could also involve developing new institutions to fill governance gaps, but this should generally be considered a secondary priority to improving coordination between, and capacities of, existing institutions due to the fragmentation and feasibility risks discussed above.

A strong regime complex for AI may sound fanciful, particularly given the need to coordinate numerous stakeholders, yet there is precedent in other areas of international policy-making, notably climate. Climate governance, like AI, is not a single policy problem and instead involves different issues such as biodiversity and carbon emissions. After decades of failure to reach meaningful global agreements, culminating in the unfruitful 2009 UN Climate Change Conference in Copenhagen, the focus of climate governance moved from developing a centralized regime to leveraging the polycentric order by encouraging multilevel action from local governments and private companies to enact innovative policies suited to their specific responsibilities.[50] Expert work from the IPCC acts as a cornerstone for informing decentralized action.

There are limitations to drawing parallels with climate policy. Notably, AI governance is more value-laden and subject to interstate competition, suggesting cooperation may prove more challenging. Nonetheless, interstate cooperation is taking place, as well as multilayered governance when interstate action stalls. More importantly, even if we anticipate the AI regime complex being imperfect, the history of climate governance indicates that an incremental approach is more likely to support successful outcomes in a multifaceted policy area than solely relying on centralized bargaining.[51]

Considering the second criterion for political legitimacy—the presence of democratic decision-making procedures—a new AI body established in a multilateral forum like the UN would have a high degree of procedural legitimacy

[48] David G. Victor, Marcel Lumkowsky and Astrid Dannenberg, 'Determining the credibility of commitments in international climate policy', *Nature Climate Change*, vol. 12, 2022, pp. 793–800, https://doi.org/10.1038/s41558-022-01454-x.

[49] Victor Galaz et al., 'Polycentric systems and interacting planetary boundaries—emerging governance of climate change–ocean acidification–marine biodiversity', *Ecological Economics*, vol. 81, 2012, pp. 21–32, https://doi.org/10.1016/j.ecolecon.2011.11.012.

[50] Martin Jänicke, 'The multi-level system of global climate governance—the model and its current state', *Environmental Policy and Governance* 27: 2, 2017, pp. 108–21, https://doi.org/10.1002/eet.1747.

[51] Kenneth W. Abbott, 'Strengthening the transnational regime complex for climate change', *Transnational Environmental Law* 3: 1, 2014, pp. 57–88, https://doi.org/10.1017/S2047102513000502.

due to being delegated authority from member states. But such an institution would not be unimpeachable. An IAEA-like body for AI would be at higher risk of regulatory capture than the original IAEA because of the commercial incentives for big tech companies to shape governance, potentially undermining the body's democratic mandate. These companies are already shifting the narrative from the harms currently caused by their products to those that are more speculative, which was successful in framing a high-profile AI Safety Summit,convened by the UK in November 2023.[52] A semi-centralized regime would help mitigate these issues, but first- and second-order cooperation problems mean that there is only a slim possibility of establishing multiple new institutions through a strong democratic procedure.

A strong regime complex would have different benefits and drawbacks with respect to procedural legitimacy. It is unlikely that many—if any—regime complex institutions would receive the same universal mandate from the UN to govern AI, weakening the democratic mandate from states. A decentralized model also risks resource-constrained civil society organizations being unable to participate at multiple forums.[53] However, a regime complex possesses a strong democratic procedure when considering input beyond the state, as decentralized governance facilitates consideration of various issues and stakeholders. It allows for a diverse range of inputs at different levels of governance, which is particularly important because of the legitimacy problem affecting international organizations in terms of citizen acceptance.[54]

Policy recommendations

Strengthening the existing AI regime complex rather than developing new centralized institution(s) is the more desirable and realistic governance option. This pathway is no panacea for addressing myriad global AI governance challenges, but it supports incremental progress across multiple layers of governance that can catalyse meaningful change. We offer recommendations to strengthen the existing AI regime complex by improving coordination and procedural legitimacy. These recommendations provide first steps in a path towards a stronger regime complex.

Coordination needs to be improved between international institutions and within the landscape more broadly. Forthcoming work by the UN High-Level Advisory Body on AI to map the international AI ecosystem and provide recommendations is an important opportunity to begin addressing this. Utilizing channels of communication and negotiations to agree on remits between international institutions will be important, but the polycentric nature of the regime complex means the highest priority is aligning nodes around common goals developed by an expert body. Authoritative expert information can support alignment

[52] 'Stop talking about tomorrow's AI doomsday when AI poses risks today'.

[53] Cihon, Maas and Kemp, 'Fragmentation and the future'.

[54] Lisa Dellmuth, Jan Aart Scholte, Jonas Tallberg and Soetkin Verhaegen, 'The elite–citizen gap in international organization legitimacy', *American Political Science Review* 116: 1, 2022, pp. 283–300, https://doi.org/10.1017/S0003055421000824.

and in turn, produce reinforcing action by different states and across multiple levels of governance. A UN body would be desirable for fulfilling this role, yet scepticism about viability is warranted for the reasons discussed above. The UK's new AI Safety Institute is another possibility, given that it has been specifically established to inform global policy-making. However, it was established unilaterally and has received a lukewarm reception from other countries, suggesting that it will face difficulties in becoming the recognized centre of expertise.

The OECD could act as a stopgap for providing expert guidance related to the economic and societal impacts of AI, given its existing work and the role the organization plays as a 'meditative' forum for constructing and disseminating research and policy ideas that subsequently shape policy at a national level: for example, by developing indicators and standard forms of measurement, as well as rankings that permit cross-country comparisons.[55] Expanding the OECD's AI work could include outputs like an economic impact ranking, frameworks for policy harmonization, indicators for good governance and recommendations for mitigating specific risks. This type of authoritative information would support evidence-based cooperation and peer pressure between states that can enable agreement. It could also inform governance efforts by subnational and private sector actors or be used by civil society organizations lobbying governments.

The key risk of relying on the OECD is a perceived lack of legitimacy on account of being a predominantly Europe-based organization, leading to the ignoring of its outputs by some states. Having specific projects delegated to the OECD from a more representative organization, like the G20, could mitigate this risk. There is a strong precedent for the OECD playing this 'palliative governance' role by using its technical expertise to grease the 'wheels of global governance' in support of other bodies, with the G20 having delegated aspects of international tax policy to the OECD serving as a notable example.[56] Scepticism is warranted as to the G20's willingness to provide such a mandate, particularly given first-order cooperation problems and China's continued emphasis on leveraging the UN for AI governance.[57] That said, the G20 previously drawing on expert work from the OECD to inform its AI principles, combined with China's increasing engagement with the OECD and its participation at the UK's AI Safety Summit, suggests these barriers are not insurmountable. Building on a proven expert body with which key stakeholders are already engaging is also more likely to yield positive results in overcoming cooperation problems than starting afresh.

Strengthening the regime complex also requires improving democratic procedures in current rule-making by scrutinizing whether existing nodes are fulfilling appropriate functions. Here, too, forthcoming work by the UN High-Level Advisory Body provides an opportunity for assessing whether the right institu-

[55] Rianne Mahon and Stephen McBride, 'Standardizing and disseminating knowledge: the role of the OECD in global governance', *European Political Science Review* 1: 1, 2009, pp. 83–101, https://doi.org/10.1017/S1755773909000058.

[56] Richard Woodward, *The Organisation for Economic Co-operation and Development (OECD)* (Abingdon and New York: Routledge, 2009).

[57] The Cyberspace Administration of China, *Global AI Governance Initiative*, 18 Oct. 2023.

tions are undertaking the right type of work. An example of where scrutiny should be exercised is in the highly value-laden work being undertaken in international standards bodies, given these bodies' membership and procedures. Suggestions have been made to increase civil society participation in these bodies to enhance their procedural legitimacy, yet there is a significant negative association between the technical complexity of regulatory proposals and the degree of 'mobilized dissent' by civil society actors, indicating that these organizations are unlikely to be able to meaningfully contribute due to resource and expertise limitations.[58] In such cases, it is unlikely that simply improving organizational procedures would be sufficient.

To promote stronger global AI governance, it is necessary to shift the discussion from focusing on which type of international AI body should be established to broader questions of how coordination and democratic procedures can be improved. It took decades of cooperation failure in climate governance to move towards a more decentralized model. We should not make the same mistake with AI.

[58] Stefano Pagliari and Kevin Young, 'The interest ecology of financial regulation: interest group plurality in the design of financial regulatory policies', *Socio-Economic Review* 14: 2, 2016, pp. 309–37, https://doi.org/10.1093/ser/mwv024.