Proposal for an International AI Governance Organisation to Prevent Catastrophic AI Outcomes

Judit BAYER[1]

DOI: 10.29180/978-615-6886-28-6_3

Abstract

In light of escalating geopolitical tension and accelerating artificial intelligence (AI) developments, this article will argue for the formation of an *International AI Governance Organisation* spearheaded by the European Union. The organisation's core mission will be the prevention of catastrophic outcomes from Artificial General Intelligence (AGI) or autonomous AI as well as rogue use of frontier AI systems including existential threats to humanity. The EU will be suggested as a natural leader given its strong regulatory legacy and global credibility. Positioning AI as a *shared security risk* – not merely a matter of innovation or competitiveness – the paper will argue for a narrow mandate focused solely on frontier risks. This mandate will aim to foster global consensus, and transcending value-laden ethical conflicts. Drawing on comparative insights from organisations such as the Internet Corporation for Assigned Names and Numbers (ICANN), the Institute of Electrical and Electronics Engineers (IEEE), the International Atomic Energy Agency (IAEA) and the World Trade Organisation (WTO), the article will lay out the institutional design of the proposed *International AI Governance Organisation* as a multistakeholder body comprising states, industry, researchers, and civil society. Equal representation will balance geopolitical power with technical and public interest input. Crucially, the paper will introduce application-programme-interface-level governance as a novel, artificial bottleneck and enforceable policy lever. In addition, the study will argue that transparency, traceability, and ethical use of frontier AI systems can be achieved via controlled application-programme-interface (API) access. In parallel, the article will explore the feasibility of creating a new ICANN-backed top-level domain (.aiapi) for licenced APIs, mirroring domain name system governance mechanisms to enhance oversight. The article will discuss that this approach could sidestep traditional enforcement weaknesses demonstrated by international law through leveraging functional chokepoints rather than sovereign coercion. It will also identify potential risks – especially organisational capture – and offer mitigation strategies through institutional checks, transparency, and civil society oversight. Finally, the study will position the proposed organisation not only as an urgent necessity but also as a strategic opportunity for the EU to exercise global leadership in shaping a cooperative AI future.

Keywords: AI governance, AI safety, global cooperation, multistakeholder, Brussels Effect

Introduction: From Mutual Suspicion to Collective Survival

The emerging global artificial intelligence (AI) landscape is defined by uncertainty, mistrust, and accelerating technological capabilities. Despite this volatile terrain, this paper argues that it is both possible and essential to establish a new international governance framework – *the International AI Governance Organisation (IAIGO)* – with the explicit aim of preventing catastrophic outcomes from AI, particularly AGI (Artificial General Intelligence) or autonomous AI systems.

[1] BGE, Kommunikáció tanszék, egyetemi docens, ORCID: 0000-0003-0558-807X, bayer.judit@uni-bge.hu

The urgency arises from the increasing fragmentation of international relations and the absence of a central authority capable of coordinating global action. The threat landscape is no longer theoretical: AI misuse, strategic disinformation, military automation, and rogue actor deployments all contribute to a real risk of existential harm. Unlike nuclear threats, AI development is decentralized and fast-moving, which makes the comparison to traditional state-centred regulatory regimes like the International Atomic Energy Agency (IAEA) only partially valid.

IAIGO is envisioned not as a comprehensive ethical framework or a replacement for a national regulation, but as a minimalist but enforceable institution with a *narrow mandate*, i.e. to prevent global catastrophe stemming from advanced AI misuse. This scope is deliberately minimal to foster participation across geopolitical divides.

Institutional Models and Lessons for IAIGO

To build IAIGO, the paper undertakes a comparative analysis of existing international organizations with varied mandates, enforcement powers, and governance models. Four institutions – the Internet Corporation for Assigned Names and Numbers (ICANN), the Institute of Electrical and Electronics Engineers (IEEE), the International Atomic Energy Agency (IAEA) and the World Trade Organisation (WTO) – were selected from a wide range of international institutions in the initial sample to represent various multistakeholder and intergovernmental models.

ICANN's management of domain names and internet infrastructure offers a precedent for technical coordination through private contracts, while IEEE demonstrates the power of consensus-based standard-setting. The IAEA draws on lessons from international monitoring and escalation procedures in high-risk domains, while the WTO illustrates how enforcement mechanisms can be grounded in treaty-backed dispute settlement systems with practical economic leverage. The analysis reveals that legitimacy, flexibility, and enforceability often trade off against one another. Those organizations that are fast-moving and adaptive tend to lack coercive powers, while those with enforcement tools frequently suffer from politicization and inefficiency. IAIGO must therefore adopt a hybrid structure that balances legitimacy from state participation with operational agility derived from technical experts, industry actors, and civil society.

Narrow Mandate, Wide Reach

The paper makes a compelling case for a minimalist mandate: IAIGO should focus solely on avoiding catastrophic outcomes from AI including AGI misuse or failure. This tightly scoped objective is essential for achieving consensus among actors with vastly different political, ethical, and economic worldviews. The mandate excludes divisive issues such as privacy, bias, or market regulation – important as they are – because global agreement on such topics remains elusive. Instead, IAIGO would limit itself to promoting transparency, traceability, and pre-deployment risk mitigation for advanced AI systems capable of causing massive harm. Such a narrowly focused goal increases the political viability of the institution while laying the groundwork for further cooperative frameworks. Over time, IAIGO could serve as a platform for developing auxiliary branches and working groups focused on ethics, human rights, or data governance. Still, the initial function of IAIGO must remain pragmatically restrained.

To ground the proposal in real-world institutional logic, the paper compares IAIGO's potential structure to four existing international bodies.

Table 1. *Classification of selected international organisations based on mandate, constituency, enforcement and normative influence*

Source: *author*

Institution	Mandate	Constituency	Enforcement	Normative Influence
ICANN	Manage Internet identifiers	Multi-stakeholder	Contract-based	High (Internet architecture)
IEEE	Technical standards	Engineers, researchers	Voluntary	High (industry-wide)
IAEA	Nuclear safety	States (UN)	Monitoring & UN escalation	High (nuclear sector)
WTO	Trade regulation	States	Binding legal system	High (global trade)

IAIGO draws on this mixed architecture, and aspires to balance the legitimacy and oversight of states with the expertise and operational capability of private actors.

Multistakeholder Composition and Governance Model

IAIGO's legitimacy would derive from its global inclusiveness and its structured balance of power between the different actor types: states, corporations, researchers, and civil society. The organisation is envisioned as a multistakeholder network with a central node and decentralized working arms – capable of rapid technical response and global norm articulation.

The inclusion of states is crucial for institutional credibility and enforcement. At the same time, corporate actors provide the technical infrastructure and deployment pathways that no intergovernmental body can access on its own. Civil society ensures normative balance and public accountability, while researchers inject frontier-relevant expertise into both rulemaking and crisis management.

Two operative substructures – a Rapid-Response Operative Body (RROB) and a Scientific Crisis Board – would enable IAIGO to act swiftly when signs of AGI misalignment, dangerous deployment, or malicious use emerge. Unlike traditional intergovernmental bodies that take months or years to reach decisions, these units could convene within hours, incorporating real-time input from national authorities and scientific advisors.

Decision-Making and Layered Agreements

The foundational charter would be limited to a brief, culturally neutral set of commitments – the IAIGO Principles – in connection with transparency, non-malicious deployment, and minimal standards for AGI risk mitigation. These principles would form the legal and normative backbone of the institution and would serve as the conditions for participation and continued recognition within the global AI ecosystem.

To support regional diversity and interoperability, IAIGO would also facilitate second-layer agreements between regional actors or industry sectors. These conventions would allow differentiated standards to emerge within broader technical compatibility frameworks. For example, the EU or ASEAN could establish more ambitious norms under the IAIGO umbrella, while remaining interoperable with global standards on traceability and accountability.

Working groups and technical committees, modelled after IEEE and ICANN, would draft and maintain evolving policy instruments in response to technological development. This two-tiered architecture ensures that IAIGO remains grounded in a minimal consensus while enabling dynamic expansion through voluntary alignment and technical integration.

Enforceability through Technical Bottleneck

Recognizing the failure of traditional international law to enforce compliance in fast-moving technological domains reliably, it was found that control over a technical bottleneck, or gatekeeping is the most secure way to ensure consistent compliance without political exceptions.

This article proposes a novel approach to effective supervision: governance through API (Application Programme Interface). Controlling API access points can serve as a novel, enforceable policy lever and an artificial bottleneck, which ensures transparency, traceability, and the ethical use of frontier AI systems. API access already serves as a natural chokepoint in the AI ecosystem. Leading companies such as OpenAI, Google, and Anthropic regulate use through contractual terms that prohibit deployments for military, surveillance, or disinformation purposes. IAIGO would extend this logic by establishing a centralized registry or licensing system for APIs that are used to access or deploy high-risk models. Violation of IAIGO principles would lead to removal from the registry, revocation of access keys, or inclusion on a public list of non-compliant actors. Because powerful models are increasingly hosted via APIs rather than being downloaded as software, this creates a real and scalable enforcement mechanism that does not require state-based coercion or legal sanctions. The limitations of governance through API are also explored: such as bypass via unregistered APIs, closed-loop deployments, cloning API, and regulatory capture. Additional limitations include ongoing AI risks posed by narrow AI systems.

To further strengthen enforceability and visibility, the paper proposes the creation of a dedicated top-level domain (.aiapi) in collaboration with ICANN. APIs used to access high-risk models would be registered under this domain and would be subject to IAIGO certification and ongoing compliance. This would mirror ICANN's approach to coordinating global internet infrastructure through domain name contracts and root server oversight. The .aiapi domain would serve both a symbolic and practical function: signalling compliance to users, enabling automated oversight tools, and offering a reputational incentive for participation. While not all AI APIs could be captured under such a scheme – particularly those used in private, closed-loop systems –, the public visibility and reputational pressure attached to the domain would be sufficient to create a new standard of transparency and auditability in AI deployment.

Role of the EU: Seeding the Initiative

The European Union is uniquely situated to launch and legitimize IAIGO. As a global regulatory actor, the EU has already demonstrated its capacity to set norms beyond its borders, most notably through the General Data Protection Regulation (GDPR) and the Digital Markets

Act and – somewhat hesitantly – by the AI Act. In the context of AI, the EU has both the institutional apparatus and the normative ambition to serve as the initiator of IAIGO. Unlike the United States, whose trust deficit in multilateral forums has grown, or China, whose governance models face ideological resistance, the EU retains the diplomatic neutrality and internal diversity to foster trust across geopolitical divides.

Crucially, the EU's role in seeding IAIGO is not about exporting its values in a wholesale manner. The proposal explicitly warns against embedding culturally divisive principles – such as privacy or content regulation – into IAIGO's founding charter. Instead, the EU would convene global actors around a minimalist mission: preventing global catastrophe. In doing so, it reinforces its role as a steward of international norms in the digital age, while safeguarding humanity's collective future.

Risk Management and Institutional Integrity

The success of IAIGO will depend not only on its technical tools and diplomatic strategies, but also on its internal integrity and resilience against capture. This paper identifies multiple capture risks: dominance by powerful state actors, entrenchment of leading AI corporations, and structural exclusion of smaller nations or civil society voices. To mitigate these issues, IAIGO must implement layered accountability. Transparency is foundational: decision-making processes, stakeholder influence, audit results, and enforcement actions must all be publicly visible and subject to independent review. The same visibility that powers reputational incentives can also check backroom deals and undue influence. Second, the organisation must be structurally pluralistic. No single actor type should dominate the decision pipeline. States must retain veto power over foundational standards but not over technical enforcement; corporations must co-regulate APIs but not unilaterally define risk categories; civil society must participate not merely as a symbolic presence but as a source of counterbalance and normativity. Third, contestability mechanisms must be built into IAIGO's operational design. These include rotating seats, open nomination of working group members, appeal rights for stakeholders, and public comment periods for key decisions. These features ensure that IAIGO evolves alongside technological and political realities rather than ossifying into a cartel of early adopters.

Conclusion: Realism Without Paralysis

The IAIGO proposal is ambitious but rooted in institutional realism. It does not call for a universal treaty, a single global regulator, or a halt to innovation. Instead, it proposes a practical, enforceable governance structure focused solely on catastrophic AI risks using multistakeholderism, technical enforcement, and minimalist consensus. The use of API access as a regulatory lever is one of the few feasible options in a fragmented international order. By building upon existing contractual ecosystems and digital chokepoints, IAIGO sidesteps the need for state coercion or treaty ratification while still creating real incentives for compliance. Ultimately, the value of IAIGO is not only in preventing catastrophe but in demonstrating that proactive, anticipatory governance is still possible in the digital age. It is a model of what global cooperation could look like: lean, legitimate, enforceable and future-facing.

References

5 Best Practices for API Governance in 2025 - API7.Ai. '5 Best Practices for API Governance in 2025 - API7.Ai'. Accessed 29 September 2025. https://api7.ai/blog/api-governance-best-practices-2025.

'AI 2027'. Accessed 24 September 2025. https://ai-2027.com/slowdown.

AIMultiple. 'When Will AGI/Singularity Happen? 8,590 Predictions Analyzed'. Accessed 24 September 2025. https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/.

Almeida, Patricia Gomes Rêgo de, Carlos Denner dos Santos, and Josivania Silva Farias. 'Artificial Intelligence Regulation: A Framework for Governance'. *Ethics and Information Technology* 23, no. 3 (2021): 505–25. https://doi.org/10.1007/s10676-021-09593-z.

Almeida, Patricia, Carlos Santos, and Josivania Silva Farias. 'Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance'. Paper presented at Hawaii International Conference on System Sciences. 2020. https://doi.org/10.24251/HICSS.2020.647.

'America's Reputation Drops across the World | Ipsos'. 17 April 2025. https://www.ipsos.com/en/americas-reputation-drops-across-the-world.

'Archive.Icann.Org/En/Structure/Structure-Map.Htm'. Accessed 24 September 2025. https://archive.icann.org/en/structure/structure-map.htm.

Batool, Amna, Didar Zowghi, and Muneera Bano. 'AI Governance: A Systematic Literature Review'. *AI and Ethics* 5, no. 3 (2025): 3265–79. https://doi.org/10.1007/s43681-024-00653-w.

Batool, Amna, Didar Zowghi, and Muneera Bano. *Responsible AI Governance: A Systematic*

Literature Review. Version 1. arXiv, 2024. https://doi.org/10.48550/ARXIV.2401.10896.

Bradford, Anu, R. Daniel Kelemen, and Tommaso Pavone. 'Europe Could Lose What Makes It

Great'. *Foreign Affairs*, 21 April 2025. https://www.foreignaffairs.com/europe/europe-could-lose-what-makes-it-great.

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

Cihon, Peter, Matthijs M. Maas, and Luke Kemp. 'Fragmentation and the Future: Investigating Architectures for International AI Governance'. *Global Policy* 11, no. 5 (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'. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA), AIES '20, Association for Computing Machinery, 7 February 2020, 228–34. https://doi.org/10.1145/3375627.3375857.

Cihon, Peter, Jonas Schuett, and Seth D. Baum. 'Corporate Governance of Artificial Intelligence in the Public Interest'. *Information* 12, no. 7 (2021): 275. https://doi.org/10.3390/info12070275.

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.

De Almeida, Patricia Gomes Rêgo, Carlos Denner Dos Santos, and Josivania Silva Farias. 'Artificial Intelligence Regulation: A Framework for Governance'. *Ethics and Information Technology* 23, no. 3 (2021): 505–25. https://doi.org/10.1007/s10676-021-09593-z.

Erdélyi, Olivia J., and Judy Goldsmith. 'Regulating Artificial Intelligence: Proposal for a Global Solution'. *Government Information Quarterly* 39, no. 4 (2022): 101748. https://doi.org/10.1016/j.giq.2022.101748.

Erdélyi, Olivia J., and Judy Goldsmith. 'Regulating Artificial Intelligence: Proposal for a Global Solution'. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, ACM, 27 December 2018, 95–101. https://doi.org/10.1145/3278721.3278731.

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

Kerry, Cameron F., Joshua P. Meltzer, Andrea Renda, and Andrew W. Wyckoff. 'Network Architecture for Global AI Policy | Brookings'. Think Tank. Brookings, 10 February 2025. https://www.brookings.edu/articles/network-architecture-for-global-ai-policy/.

Khalid, Asma. 'Biden and Xi Take a First Step to Limit AI and Nuclear Decisions at Their Last Meeting'. Politics. *NPR*, 16 November 2024. https://www.npr.org/2024/11/16/nx-s1-5193893/xi-trump-biden-ai-export-controls-tariffs.

Koniakou, Vasiliki. 'From the "Rush to Ethics" to the "Race for Governance" in Artificial Intelligence'. *Information Systems Frontiers* 25, no. 1 (2023): 71–102. https://doi.org/10.1007/s10796-022-10300-6.

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

'Network Architecture for Global AI Policy | Brookings'. Accessed 29 September 2025. https://www.brookings.edu/articles/network-architecture-for-global-ai-policy/.

Pasquale, Frank. 'From Territorial to Functional Sovereignty: The Case of Amazon'. *LPE Project*, 6 December 2017. https://lpeproject.org/blog/from-territorial-to-functional-sovereignty-the-case-of-amazon/.

'Root Zone Management'. Accessed 29 September 2025. https://www.iana.org/domains/root.

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 (2022): 815–45. https://doi.org/10.1007/s11558-021-09445-x.

Taeihagh, Araz. ‘Governance of Artificial Intelligence’. *Policy and Society* 40, no. 2 (2021): 137–57. https://doi.org/10.1080/14494035.2021.1928377.

Tegmark, Max. *Life 3.0*. Unabridged. Random House, Inc., 2017.

Veale, Michael, Kira Matus, and Robert Gorwa. ‘AI and Global Governance: Modalities, Rationales, Tensions’. *Annual Review of Law and Social Science* 19, no. Volume 19, 2023 (2023): 255–75. https://doi.org/10.1146/annurev-lawsocsci-020223-040749.

Villalobos, Matthijs Maas, José Jaime. *International AI Institutions A Literature Review of Models, Examples, and Proposals*. 2023. https://law-ai.org/international-ai-institutions/.

Wallach, Wendell, and Gary E Marchant. *An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics*. n.d.

‘Why Metadata Maturity Matters for AI-Ready Data | Key Insights from Gartner | Alation’. Accessed 29 September 2025. https://www.alation.com/blog/metadata-maturity-ai-ready-data-gartner/.

Zaidan, Esmat, and Imad Antoine Ibrahim. ‘AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective’. *Humanities and Social Sciences Communications* 11, no. 1 (2024): 1121. https://doi.org/10.1057/s41599-024-03560-x.