Centre for International
Governance Innovation

Working Paper

Developing Safety Standards for an International Agreement on Advanced AI

Christo Hall

Fall 2025 term

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

This paper was completed during a co-op placement with the Global AI Risks Initiative at the Centre for International Governance Innovation and published through the organization's Digital Policy Hub program.

Copyright © 2026 by Christo Hall

The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Centre for International Governance Innovation or its Board of Directors.

Centre for International Governance Innovation and CIGI are registered trademarks.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
cigionline.org

Introduction

Many of the most challenging risks that stem from the development and deployment of cutting-edge artificial intelligence (AI) are global in nature. Powerful systems could enable rogue actors to develop chemical, biological, radiological and nuclear weapons or augment cyberattacks and disinformation campaigns (Wagman and Hubbard 2025), and misaligned systems have the potential to erroneously or intentionally make decisions that threaten human lives (Mitre and Predd 2025). In addition, mass AI adoption poses economic and labour disruption (Occhipinti et al. 2025) and power and wealth concentration (United Nations Conference on Trade and Development 2025) that could fundamentally alter the global economy and humans' relationships with work and one another. These are all cross-border risks that have the capability to affect people outside of the state in which the threat originates or where the system is developed or deployed.

At the governance level, competitive dynamics among states for pieces of the AI pie are leading to race-like conditions and a desire for technological independence (Larsen 2022). This creates the potential for a single dominant player to surface with limited international collaboration or a multipolar world where regional rules and norms could lead to fragmented governance, among other scenarios (Samson, Kalash and Zivkovic 2025). The Centre for International Governance Innovation's Global AI Risks Initiative has responded to these threats via three challenges posed to the global community:

- to realize and share global benefits;
- to mitigate global risks; and
- to ensure inclusive decision making.

Thus far, international governance of AI has been slow to respond (Roberts et al. 2024), and if the development and deployment of AI continues to accelerate, an international agreement may be essential to overcome these challenges (Cass-Beggs et al. 2024). Therefore, the purpose of the Global AI Risks Initiative is to develop the building blocks of an international agreement on advanced AI to inform future governance, if the conditions for it are ripe.

One of the building blocks of the initiative's proposal is the inclusion of technical and safety standards so that the agreement can operationalize its requirements and principles and so that developers and deployers can adhere to and confer compliance with the agreement. In the proposal, standards would also be the basis of a licensing regime that grants AI organizations the legal permission to operate a high-risk system. This working paper supports that effort and the international AI governance literature by proposing a framework for developing standards within an international agreement on advanced AI.

Who Is Responsible for Standardization?

International agreements that have utilized standards have broadly done so in one of three ways:

- **Creation:** The agreement's parties have established an internal agency, or utilized an agency created under a related agreement, that is responsible for the development of standards (for example, the Chicago Convention's establishment of the International Civil Aviation Organization [2016], or the International Atomic Energy Agency's facilitation of standards development for the Convention on Nuclear Safety).[1]
- **Reference or delegation:** The agreement sets requirements and high-level principles, and then either refers to standards already developed by an external standards development organization (SDO) or sets up a panel to assess externally developed standards' harmonization with those requirements and principles (for example, the Montreal Protocol on Substances that Deplete the Ozone Layer[2]).
- **Commission:** The European Union's "new approach" (European Commission 2002) is an example of this policy innovation that has yet to be reproduced at the international level. The new approach constitutes the European Union appointing a private SDO to develop standards on its behalf (for example, the commissioning of the European Committee for Standardization [CEN] and the European Committee for Electrotechnical Standardization [CENELEC] to produce standards within the EU AI Act). Compliance with those standards confers a presumption of conformity with the law.

Historically, standards for international agreements on topics of critical public safety have been developed by an associated agency (Albisinni 2016). However, the increasing number of standards produced by private SDOs has raised concerns about competing standards' compliance burdens and forum shopping. Coupled with governments' pragmatism to work with industry in developing rules and standards, the trend has moved away from "create" and toward "refer or delegate" (Berman 2017).

This working paper argues that in order for a future international agreement on advanced AI among states to be effective, it must possess the capacity to create standards for the following three reasons:

- There are significant gaps in advanced AI safety and governance research (Strauss et al. 2025; Apollo Research 2024), and it is unclear whether research will have matured sufficiently for standardization before the establishment of an international agreement.
- Private SDOs' processes have several shortcomings, including that they lack socio-technical and regionally diverse expertise (Dunietz et al. 2024); they are

1 See www.iaea.org/resources/safety-standards.

2 *Montreal Protocol on Substances that Deplete the Ozone Layer*, 16 September 1987, 1522 UNTS 3, 26 ILM 1550 (entered into force 1 January 1989), online: <https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-2-a&chapter=27&clang=_en>.

unrepresentative of public interests (Lenoir, Galissaire and Lucas 2025); and they facilitate competition among standards (Roberts et al. 2024).

- In a rapidly moving field, established standards-setting processes can take many years and will be too slow for adequate anticipatory governance (Roberts and Ziosi 2025).

Standards supporting an international agreement serve to operationalize its requirements and principles. They also add crucial agility to respond to a changing threat environment, provide opportunities to involve a diverse range of experts in the standardization process, and enhance legitimacy that derives from both of these benefits. However, it is not sufficient that a standards' creation function alone remains agile. The process's design should also guarantee participation from appropriate socio-technical expertise and include globally representative voices, or at least voices that represent the breadth of the agreement's signatories. An inclusive standardization process would allow advanced AI to be governed in a truly anticipatory manner (Organisation for Economic Co-operation and Development 2024).

However, if standards creation were purely the responsibility of an agency within the agreement, a safety-critical and urgent standard might not be created or amended in time for it to be operational. That would be especially true in areas where standards might have been published by a private or public SDO or consortia such as the Frontier Model Forum or the International Network of AI Safety Institutes (AISI). An adequately agile process would need to balance the creation of standards — where none exist, or where existing standards and their development processes are inadequate — with the ability for the agency to refer to sufficiently robust existing standards developed elsewhere.

An Agile Standards-Setting Framework

For standards to be developed within an agreement, the parties would need to mandate a newly appointed agency whose function would be to establish standards specifying the practices that AI providers must follow to meet the agreement's requirements. The agency would consist of a core panel of appointed experts who set the agenda for standardization and specification, similar to the way in which the executive committee of Codex Alimentarius assesses standardization gaps and proposals for new standards on behalf of the member states that make up the Codex Alimentarius Commission.[3]

In this framework, the distinction between standards and specifications is crucial to embed the requisite agility. Standardization resembles traditional standards development processes, which seek a diverse range of opinions in a consensus-led manner. It may take a few years to arrive at a publishable standard. Specification is much more flexible; its result will not be as inclusive or as legitimate but it will produce a publishable finding within months. This proposal's articulation of a specification process is modelled on elements from several national standards initiatives in Canada

3 See www.fao.org/fao-who-codexalimentarius/committees/executive-committee/about/en/.

and the United Kingdom (Standards Council of Canada 2021; British Standards Institution 2020).[4] Article 41 of the EU AI Act also provides for the development of common specifications by the act's advisory forum.[5] However, specifications in this context are intended to fill gaps where standardization development falls short, rather than serving as primary instruments of the act.[6]

The content of standards and specifications should be developed by working groups or committees but it must be consistent with the agreement's high-level principles and requirements, similar to the way in which standards developed by CEN-CENELEC on behalf of the European Commission must align with the essential requirements set out in the EU AI Act (Soler Garrido et al. 2024). Later, parties may submit requests to the conference of the parties (COP) to propose additional standards or amendments to existing ones.

Standardization Framework

The agency's panel of experts should be tasked with scope-setting and project management for each standard. The standardization process does not need to differ dramatically from established processes by private SDOs, with the exception that working groups must be sufficiently diverse and processes must be designed in ways that supress industry tactics (Corporate Europe Observatory 2025), preventing any stakeholder group from having outsized influence.

The working groups should be made up of stakeholders from industry, academia, civil society and governments, and should be representative of global interests or the diversity of the agreement's members. The typical number of active participants in AI standardization is not well discussed in the literature, but anecdotal reports indicate that international private SDOs regularly have 75 to more than 150 members who develop and deliberate on AI standards (Koene, Dowthwaite and Seth 2018; International Organization for Standardization/International Electrotechnical Commission [ISO/IEC] Joint Technical Committee 1 2022). For an inclusive international agency such as the one proposed in this working paper, the number of participants will likely be at the upper end of that range.

Parties should be able to recommend working group members to the expert panel, but the panel should have responsibility for the selection process. They would also have responsibility for selecting five to 10 members of a working group who make up an editorial team that reviews drafts at stages prior to publication. To ensure greater inclusivity among participants who cannot commit resources to an iterative standards development process, the first draft of the standard should be published online for an eight-week public consultation period. Each standard would be expected to take two to three years to complete, with annual post-publication reviews coinciding with COPs.

4 See https://scc-ccn.ca/standards/flexible-standards-based-solutions/workshop-agreement.

5 EC, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)* [2024] OJ, L 2024/1689, art 67, online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>.

6 *Ibid*, art 41.

Specification Framework

A specification process can be significantly streamlined with the goal of producing a published document within six months. It would not be as inclusive as the standardization process, but given the goal of producing an urgently needed standard, a specification provides a provisional measure while either the standardization process gets under way or the science matures in that area to a level that warrants full standardization. The first draft of a specification is not a collective effort. Instead, it is sped up by being developed only by members of the expert panel, similar to how the National Institute of Standards and Technology's (NIST's) new "Zero Drafts" pilot project seeks to speed up its AI standardization processes.[7]

A working group of around 10–15 members is then formed. Once established, this group could appoint a further 50–150 reviewers with commenting privileges only. Again, the working group members are selected by the expert panel and can be suggested by parties. To enhance inclusivity, the first draft of a specification could be published online for a four-week public consultation period. It would be the working group's responsibility to incorporate the suggestions of the reviewers, but consensus only needs to be found among working group members. A member of the expert panel would act as project manager for the specification with the responsibility of ensuring that it is published within six months.

It is recommended that specifications be formally reviewed every six months by the expert panel, which invites input from the parties at annual COPs. For topics that are judged to require and be ready for full standardization, specifications are replaced by standards once they are developed. For topics where scientific work is required before standardization can be achieved, specifications may last several years, depending on scientific progress in that area. At each review process, a determination of whether standardization can begin should be made. Outside of COPs, a process of immediate review could be triggered by an expert panel member or party if an incident, significant advancement in research or shift in research paradigms motivates it.

Designing for Uncertainty

At the time of writing, AI development, domestic regulatory appetite, geopolitics, standardization, and the maturity of AI safety science and practice are uncertain, or chaotic in some cases, and could be subject to sudden paradigm or attitude shifts. Therefore, any proposal for an international agreement's standards development process must be sensitive to the emergence of radically different scenarios (Cass-Beggs and da Mota 2026).

Many of advanced AI's global risks are due to its dual-use nature (Elsner, Atkinson and Zahidi 2025). Therefore, as capabilities increase, so do risks. With uncertainty about the pace of AI development, it is hard to predict when or if these risks might emerge. Should AI development accelerate rapidly over the coming few years due to increased investment and adoption, scaling, algorithmic breakthroughs, reinforcement learning, and/or research paradigm shifts, there is a significant chance that development will outpace the science of safety (Bengio et al. 2024). In that scenario, optimal risk management practices would not have been established, and robust standards may need to be developed rapidly.

7 See www.nist.gov/artificial-intelligence/ai-research/nists-ai-standards-zero-drafts-pilot-project-accelerate.

Alternatively, if AI development is only incremental and takes decades to pose global and severe adverse outcomes, then the science of safety may be able to keep up with development and robust standards might be available. In such a case, if those standards satisfy the conditions of the agreement, a mechanism to adopt or adapt them is desirable to prevent redundancy and multiple compliance requirements.

To address this uncertainty, this working paper adapts a framework established in a 2024 NIST report (Dunietz et al. 2024). It is proposed here as a framework for the operation of the international agreement's agency, detailing how the expert panel's decision on a topic's readiness for standardization triggers actions, including the potential development of working groups if standardization or specification is justified (see Table 1).

Table 1: Readiness-to-Standardize Framework

Level of Maturity	Example Standard	Actions Triggered
Required and already standardized	Concepts and terminology (ISO/IEC 22989:2022)	Review process as to whether the standard can be adopted or undergo amendments through specification or standardization development before adoption
Urgently needed and ready for standardization	Transparency standard detailing incident disclosure, governance structure, financial incentives and other reporting	Specification development begins and initiation of information-gathering phase of standardization process
Needed but requires more scientific work	Testing and evaluating data sets	Specification development begins and submission of research recommendations to collaborative partners such as the AISI Network or the UN Independent International Scientific Panel on AI
Needed but requires significant scientific work	Interpretability and explainability techniques	Submission of research recommendations to collaborative partners

Source: Author.

Recommendations

To realize and share global benefits, mitigate global risks and ensure inclusive decision making for advanced AI, international cooperation (Roberts et al. 2024) and potentially only sufficiently agile and anticipatory international hard law is required (The Elders 2024). To accomplish this, an international agreement on advanced AI should incorporate standards that confer compliance with its requirements and principles and could be the basis of its licensing regime. However, public and private standardization processes are flawed in several ways and for an agile standards-setting framework to be devised, it must:

- have the capacity to both create standards from scratch and refer to suitable standards developed elsewhere;
- be sufficiently inclusive and ensure that no stakeholder group, interest or region exerts outsized influence over the process and its outcomes;
- design flexible standards products that can be produced rapidly; and

- develop a continuous monitoring mechanism that iteratively reviews and refreshes standards.

There are several research avenues and areas where innovation could support such a framework and help lay the foundations for an international agreement on advanced AI. These include:

- conducting research in several areas of risk management so that safety standards are based on scientific consensus;
- defining appropriate fora for developing advanced AI standards and preventing competition between standards;
- developing robust verification methods and conformity assessment frameworks to support licensing regimes for high-risk systems; and
- formulating strategies that advocate for and champion global fora for multilateral governance.

Ultimately, geopolitical power struggles, especially those that stem from strategic competition over AI (Esposito 2025), will need to abate before an effective international agreement can be reached. With considerable uncertainty about the pace of AI development, continued work is required to build the capacity and establish the conditions for robust risk management and international governance. As governments stimulate investment in and promote the rapid adoption of AI to create value in their economies, they must also ensure that breakneck development and deployment neither prevent the establishment of effective safety practices nor outpace the ability of standards to diffuse them.

About the Author

Christo Hall is a senior policy adviser at SickKids' Child Health Policy Accelerator, where he advocates for Canadian federal child-focused online safety legislation. Previously, he was a policy researcher with the Centre for International Governance Innovation's Global AI Risks Initiative and McMaster University's Digital Society Lab. He also spent more than a decade working in publishing and life sciences/biotech communications.

Works Cited

Albisinni, Francesco Giovanni. 2016. “The Rise of Global Standards: ICAO’s Standards and Recommended Practices.” *Italian Journal of Public Law* 8 (1): 203–31. www.ijpl.eu/wp-content/uploads/2022/10/13.Albisinni.pdf.

Apollo Research. 2024. “The Evals Gap.” *Apollo Research* (blog), November 11. www.apolloresearch.ai/blog/evalsgap.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari et al. 2024. “Managing extreme AI risks amid rapid progress.” *Science* 384 (6698): 842–45. https://doi.org/10.1126/science.adn0117.

Berman, Ayelet. 2017. “Industry, Regulatory Capture and Transnational Standard Setting.” *American Journal of International Law Unbound* 111: 112–18. https://doi.org/10.1017/aju.2017.29.

British Standards Institution. 2020. “New BSI Flex Standards enable dynamic consensus.” Press release, September 16. www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2020/september/new-bsi-flex-standards-enable-dynamic-consensus/.

Cass-Beggs, Duncan, Stephen Clare, Dawn Dimowo and Zaheed Kara. 2024. “Framework Convention on Global AI Challenges.” CIGI Discussion Paper. Waterloo, ON: CIGI. www.cigionline.org/static/documents/AI-challenges_OW6rTMD.pdf.

Cass-Beggs, Duncan and Matthew da Mota. 2026. *AI National Security Scenarios*. Summary Report from the Artificial Intelligence National Security Scenarios Workshop. Waterloo, ON: CIGI. www.cigionline.org/static/documents/AI_National_Security.pdf.

Corporate Europe Observatory. 2025. *Bias Baked In: How Big Tech Sets Its Own AI Standards*. Corporate Europe Observatory, January 9. https://corporateeurope.org/en/2025/01/bias-baked.

Dunietz, Jesse, Elham Tabassi, Mark Latonero and Kamie Roberts. 2024. *A Plan for Global Engagement on AI Standards*. NIST Trustworthy and Responsible AI. Report No. 100-5. July. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.AI.100-5.

Elsner, Mark, Grace Atkinson and Saadia Zahidi. 2025. *The Global Risks Report 2025.* 20th Edition Insight Report. January. Geneva, Switzerland: World Economic Forum. https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf.

Esposito, Mark. 2025. “AI geopolitics and data centres in the age of technological rivalry.” World Economic Forum, July 24. www.weforum.org/stories/2025/07/ai-geopolitics-data-centres-technological-rivalry/.

European Commission. 2002. *Methods of referencing standards in legislation with an emphasis on European legislation*. Enterprise Guides. Brussels, Belgium: European Commission. https://ec.europa.eu/docsroom/documents/3276/attachments/1/translations.

International Civil Aviation Organization. 2016. “SARPs — Standards and Recommended Practices.” In *Annex 19 to the Convention on International Civil Aviation — Safety Management*. 2nd ed. July. Montreal, QC: International Civil Aviation Organization. https://studylib.net/doc/28186186/annex-19-safety-management-2ed.

ISO/IEC. 2022. *SC 42 Strategic Business Plan Artificial Intelligence*. No. 15992. September 27. https://jtc1info.org/wp-content/uploads/2023/02/ISO-IEC_JTC1_N15992_SC42_Business_Plan2022.pdf.

Koene, Ansgar, Liz Dowthwaite and Suchana Seth. 2018. “IEEE P7003™ standard for algorithmic bias considerations: work in progress paper.” *FairWare ’18: Proceedings of the International Workshop on Software Fairness*, 38–41. https://doi.org/10.1145/3194770.3194773.

Larsen, Benjamin Cedric. 2022. “The geopolitics of AI and the rise of digital sovereignty.” Brookings, December 8. www.brookings.edu/articles/the-geopolitics-of-ai-and-the-rise-of-digital-sovereignty/.

Lenoir, Théophile, Jessica Galissaire and Jean-François Lucas. 2025. *AI Governance: Empowering Civil Society*. Final Report of the AI Dialogues Series. February. Paris, France: Renaissance Numérique. www.renaissancenumerique.org/en/publications/ai-governance-empowering-civil-society/.

Mitre, Jim and Joel B. Predd. 2025. “Artificial General Intelligence’s Five Hard National Security Problems.” Februrary. Santa Monica, CA: RAND Corporation. www.rand.org/pubs/perspectives/PEA3691-4.html.

Occhipinti, Jo-An, William Hynes, Ante Prodan, Harris Eyre, Roy Green, Sharan Burrow, Marcel Tanner et al. 2025. “Generative AI may create a socioeconomic tipping point through labour displacement.” *Scientific Reports* 15, 26050. https://doi.org/10.1038/s41598-025-08498-x.

Organisation for Economic Co-operation and Development. 2024. “Framework for Anticipatory Governance of Emerging Technologies.” OECD Science, Technology and Industry Policy Paper No. 165. April. Paris, France: OECD Publishing. https://doi.org/10.1787/0248ead5-en.

Roberts, Huw, Emmie Hine, Mariarosaria Taddeo and Luciano Floridi. 2024. “Global AI governance: barriers and pathways forward.” *International Affairs* 100 (3): 1275–286. https://doi.org/10.1093/ia/iiae073.

Roberts, Huw and Marta Ziosi. 2025. “Can we standardise the frontier of AI?” AI Governance Initiative, June 9. Oxford, UK: Oxford Martin School. https://aigi.ox.ac.uk/publications/can-we-standardise-the-frontier-of-ai/.

Samson, Paul, S. Yash Kalash and Nikolina Zivkovic. 2025. *AI-Driven Productivity Scenarios*. Special Report. Waterloo, ON: CIGI. www.cigionline.org/publications/ai-driven-productivity-scenarios/.

Soler Garrido, Josep, Sarah De Nigris, Elias Bassani, Ignacio Sanchez, Tatjana Evas, Antoine-Alexandre André and Thierry Boulangé. 2024. “Harmonised Standards for the European AI Act.” Science for Policy Brief. JRC No. 139430. European Commission. https://publications.jrc.ec.europa.eu/repository/handle/JRC139430.

Standards Council of Canada. 2021. *Canadian Standards Development: Publicly Available Specifications*. July 20. Ottawa, ON: Standards Council of Canada. https://scc-ccn.ca/resources/publications/publicly-available-specifications.

Strauss, Ilan, Isobel Moure, Tim O’Reilly and Sruly Rosenblat. 2025. “Real-World Gaps in AI Governance Research.” Preprint, arXiv, May 5. https://doi.org/10.48550/arXiv.2505.00174.

The Elders. 2024. “The Elders call for strong international leadership on AI governance after Summit of the Future.” The Elders, October 21. https://theelders.org/news/elders-call-strong-international-leadership-ai-governance-after-summit-future.

United Nations Conference on Trade and Development. 2025. *2025 Technology and Innovation Report: Inclusive Artificial Intelligence for Development*. March. Geneva, Switzerland: United Nations. https://unctad.org/system/files/official-document/tir2025_en.pdf.

Wagman, Shlomit and Sarah Hubbard. 2025. “Weaponized AI: A New Era of Threats and How We Can Counter It.” Ash Center for Democratic Governance and Innovation, April 8. https://ash.harvard.edu/articles/weaponized-ai-a-new-era-of-threats/.