CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

安远 AI
CONCORDIA AI

AI GOVERNANCE
INITIATIVE

OXFORD
MARTIN
SCHOOL

UNIVERSITY OF
OXFORD

Examining AI Safety as a Global Public Good: Implications, Challenges, and Research Priorities

Co-authors

Kayla Blomquist*, Oxford China Policy Lab, Oxford Martin AI Governance Initiative

Elisabeth Siegel*, Oxford China Policy Lab, Oxford Martin AI Governance Initiative

Ben Harack, Oxford Martin AI Governance Initiative

Kwan Yee Ng, Concordia AI

Tom David, General-Purpose AI Policy Lab

Brian Tse, Concordia AI

Charles Martinet, Oxford Martin AI Governance Initiative, Centre pour la Sécurité de l'IA

Matt Sheehan, Carnegie Endowment for International Peace

Scott Singer, Carnegie Endowment for International Peace, Oxford China Policy Lab, Oxford Martin AI Governance Initiative

Imane Bello, Future of Life Institute

Zakariyau Yusuf, Tech Governance Project

Robert Trager, Oxford Martin AI Governance Initiative

Fadi Salem, Policy Research Department, Mohammed bin Rashid School of Government

Seán Ó hÉigeartaigh, AI: Futures and Responsibility Programme, University of Cambridge

Jing Zhao, School of Public Policy and Management, Tsinghua University

Kai Jia, School of International and Public Affairs, Shanghai Jiao Tong University

* Equal contribution. Name order randomized.

Given the large number of authors, authorship does not imply agreement with every point made in this paper.

Acknowledgments

The authors thank Irene Solaiman, Mohan Kankanhalli, and Nate Sharadin for their thoughtful review and constructive suggestions on this report.

Contents

Executive Summary

As artificial intelligence (AI) systems become more powerful and integrated into daily life and global infrastructure, ensuring their safe development and deployment has emerged as one of the most pressing governance challenges of our time. While current narrow AI systems already have significant impacts in specific domains, advanced AI systems could fundamentally transform life through their potential for recursive self-improvement and general problem-solving capabilities, making their development and governance a uniquely critical challenge for humanity's future.

Drawing on lessons from climate change, nuclear safety, and global health governance, this analysis examines **whether and how applying the framework of a "public good" could help us better understand and address the challenges posed by advanced AI systems**. A "public good" is a commodity that is available to all and that can be used without reducing its availability to others.

This paper analyzes global public good literature frameworks and emerging AI governance challenges. Our analysis reveals several key challenges for overcoming coordination problems:

1. **Balancing Collective Responsibility with Targeted Accountability**: AI safety requires broad cooperation, but this must not diminish the accountability of leading AI developers and states that possess disproportionate power and leverage to ensure safe development. However, the stark disparity between nations developing frontier AI versus nations primarily implementing AI created elsewhere has fomented complex dynamics for international cooperation, such as by constraining global cooperation on safety measures to a few key decision-makers.

2. **Safety-Capability Entanglement**: Pursuing some critical AI safety measures may also involve advancing capabilities; alternatively, some critical AI safety measures may even require advanced capabilities to implement. These aspects create tension between the goals of sharing safety advances (on the one hand) and limiting the spread of AI capabilities with risky security or military implications (on the other).

3. **Development Equity**: It is important to ensure that AI safety requirements do not unduly constrain AI's involvement in strategies to achieve global and sustainable development goals like poverty reduction and also do not perpetuate long-standing inequities in the global system.

Rather than advocating for specific policy measures, this analysis advances our understanding of how collective action mechanisms might effectively address AI safety challenges while promoting equity and maintaining clear lines of responsibility. In addition, we offer avenues for further research in studying the application of the global public goods framework to AI safety.

Introduction

Rapid advancement of AI capabilities has sparked both enthusiasm for its potential benefits and concern about emerging collective challenges that transcend traditional boundaries.[1] While AI technologies could help address global challenges in areas such as healthcare, climate change, and economic development,[2] they also present risks, ranging from contemporary concerns about algorithmic bias[3] and privacy to fundamental questions about autonomy and safety.[4] These opportunities and challenges may manifest differently across regions and communities, as some nations push the frontier of AI development while others struggle to build fundamental infrastructure for basic AI adoption and development. Thus, AI presents complex governance

[1] See, for example: Hruby, Jill, and M. Nina Miller. "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems." Nuclear Threat Initiative, 2021. https://www.jstor.org/stable/resrep40076. Moore, Phoebe V. "Osh and the Future of Work: Benefits and Risks of Artificial Intelligence Tools in Workplaces." In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body and Motion*, edited by Vincent G. Duffy, 11581:292–315. Cham: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-22216-1_22.'; O'Mathúna, Dónal, and Ron Iphofen. "Automated Justice: Issues, Benefits and Risks in the Use of Artificial Intelligence and Its Algorithms in Access to Justice and Law Enforcement." In *Ethics, Integrity and Policymaking: The Value of the Case Study*. Research Ethics Forum Ser, v. 9. Cham: Springer International Publishing AG, 2022.

[2] Miailhe, N., C. Hodes, A. Jain, N. Iliadis, S. Alanoca, and J. Png. "AI for Sustainable Development Goals." Delphi – Interdisciplinary Review of Emerging Technologies 2, no. 4 (2019): 207–16. https://doi.org/10.21552/delphi/2019/4/10.

[3] Kordzadeh, Nima, and Maryam Ghasemaghaei. "Algorithmic Bias: Review, Synthesis, and Future Research Directions." European Journal of Information Systems 31, no. 3 (May 4, 2022): 388–409. https://doi.org/10.1080/0960085X.2021.1927212.

[4] Curzon, James, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib. "Privacy and Artificial Intelligence." IEEE Transactions on Artificial Intelligence 2, no. 2 (April 2021): 96–108. https://doi.org/10.1109/TAI.2021.3088084.

challenges at local, regional, and global scales. The intersection of rapid technological advancement and the need for collective action raises questions about whether and how different stakeholders—from major AI powers to emerging economies—can coordinate approaches to AI development, deployment, and risk management. Further, this disparity in AI capabilities and resources highlights the need for approaches that balance global development needs with frontier safety considerations.[5]

Recent international dialogues have explored frameworks of collective action and public goods as approaches to address global challenges arising from AI advancement from a variety of perspectives (see Appendix A), signaling growing recognition that this concept as applied to AI deserves serious consideration. These discussions examine how different aspects of AI development, deployment, and safety might be understood as public goods at various scales—from local to global. The public goods lens illuminates both the underinvestment in critical areas such as safety research and infrastructure development, and the challenges in coordinating action across diverse stakeholders with varying capabilities and priorities. Framing AI safety as a global public good would imply that knowledge, measures, and practices that ensure the safety of AI systems be universally accessible, non-excludable, and beneficial to all, regardless of individual contributions or geographical boundaries.

In discussions of AI and governance across the world to date, numerous potential framings have been discussed. Some suggest that AI technologies themselves might constitute global public goods, particularly in their potential to address collective challenges in areas such as climate change, public health, and sustainable development.[6] This framing emphasizes how AI capabilities, when developed and deployed equitably, could provide benefits to humanity that are non-rivalrous (one country's possession would not reduce another country's possession) and non-excludable (one country cannot prevent other countries from possessing it), through enhanced problem-solving capabilities, improved resource allocation, and accelerated scientific discovery. However, recent policy shifts, such as the Biden administration's executive orders in January 2025 toward restricting international AI technology transfer, may challenge this view of AI as a global public good, as emerging protectionist measures effectively create exclusion mechanisms that could divide the world into AI "haves" and "have-nots."[7] Relatedly, the increased capability levels and proliferation of AI systems have caused concern about the collective challenges posed by their development and deployment.[8] Some of these challenges, such as algorithmic bias and privacy, are already found in contemporary AI systems, whereas longer-term questions concern how to maintain human agency and prevent catastrophic risks.[9]

[5] Adan, S. N., and Trager, R.F., et al (2024) "Voice and Access in AI: Global AI Majority Participation in Artificial Intelligence Development and Governance, Oxford Martin AI Governance Initiative," Oxford Martin School AI Governance Initiative White Paper.

[6] Truby, Jon. "Governing Artificial Intelligence to Benefit the UN Sustainable Development Goals." Sustainable Development 28, no. 4 (July 2020): 946–59. https://doi.org/10.1002/sd.2048.

[7] The White House. "Executive Order on Advancing United States Leadership in Artificial Intelligence Infrastructure." The White House, January 14, 2025. https://www.whitehouse.gov/briefing-room/presidential-actions/2025/01/14/executive-order-on-advancing-united-states-leadership-in-artificial-intelligence-infrastructure/.

[8] Chowdhury, Rumman. "AI Desperately Needs Global Oversight." Wired. Accessed January 2, 2025. https://www.wired.com/story/ai-desperately-needs-global-oversight/.

[9] Bengio, Yoshua et al., Managing extreme AI risks amid rapid progress. *Science* 384, 842–845 (2024).

Thus, ensuring the safe development of advanced AI systems—that is, protecting against AI systems' negative externalities while preserving their benefits—might itself constitute a critical global public good.

This paper examines how public goods frameworks might inform the management of AI-related risks and externalities, analyzing both the theoretical foundations of the global public goods framework and its practical implications. The analysis considers questions of responsibility and equity, exploring how such frameworks might help address disparities in AI development while ensuring broad participation in safety efforts. Rather than advocating for specific policy measures, we seek to develop a research agenda to inform future governance efforts and international cooperation.

"Global Public Goods" as Foundational Frameworks for International Progress

The concept of a "global public good" (GPG) has emerged as a powerful analytical framework in modern economics and policy discussions, offering structured approaches to addressing collective challenges that transcend national boundaries.[10] Understanding the potential application of this framework to a world with advanced AI requires examining its core characteristics and practical implications for governance and coordination.

Global public goods differ from traditional public goods by providing benefits on a worldwide scale, but exhibit the same basic characteristics.[11] The essential distinction of a "public good" is shown in the following matrix:[12]

Table 1. Categorizing Types of Goods

	Rivalrous	**Non-Rivalrous**
Excludable	Private Good	Club Good
Non-Excludable	Common Good	Public Good

DOI:10.1126/science.adn0117; Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete Problems in AI Safety." arXiv, 2016. https://doi.org/10.48550/ARXIV.1606.06565.

[10] Key works include: Paul A. Samuelson, "The Pure Theory of Public Expenditure," *Review of Economics and Statistics* 36 (November 1954): 387–89; and idem, "A Diagrammatic Exposition of a Theory of Public Expenditure," *Review of Economics and Statistics* 37 (November 1955): 350–56.
Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Harvard University Press, 1965. https://doi.org/10.4159/9780674041660.
Oakland, William H. "Chapter 9 Theory of Public Goods." In the *Handbook of Public Economics*, 2:485–535. Elsevier, 1987. https://doi.org/10.1016/S1573-4420(87)80004-6.

[11] Olson, Mancur. The Logic of Collective Action: Public Goods and the Theory of Groups, With a New Preface and Appendix. Harvard University Press, 1965. https://doi.org/10.4159/9780674041660.

[12] https://www.tandfonline.com/doi/full/10.1080/17487870.2023.2280969

This matrix highlights the two aforementioned non-excludable and non-rivalrous characteristics of global public goods, noting their differences from other kinds of goods. A *rivalrous* and *excludable* global good would be a private good, such as a cross-border oil reserve; an *excludable* and *non-rivalrous* good would be a club good, such as satellite networks or academic journals; and a *non-excludable* and *rivalrous* good would be a common good, such as ocean fish stocks or Antarctic resources. Meanwhile, examples of a *non-rivalrous*, *non-excludable* and thus *public* good in the international context would include a stable climate, disease eradication, and a low risk of world-scale war—benefits that affect everyone globally and whose enjoyment by one party does not diminish their availability to others.[13]

Framing certain goods and services as "global public goods" carries important implications for how they should be funded, governed, and distributed equitably around the world.[14]

While goods can be public at multiple levels of analysis—including within a community, city, country, or region—this paper focuses on the global level due to AI's expected worldwide impact and the associated global coordination challenges. Markets may under-provide AI safety because its benefits are a public good that individual companies cannot fully capture, while the costs of safety measures are private and directly impact their bottom line. The underprovision challenge stems from classic free-rider dynamics observed in other global public goods contexts. Individual actors—whether nations, companies, or research institutions—may underinvest in safety measures, knowing that they can benefit from others' safety investments without bearing the costs. This dynamic may be particularly concerning for AI safety given the global scope of potential harms from inadequate safety measures. When multiple actors adopt this approach, the collective investment in safety falls below socially optimal levels, potentially leading to insufficient protection against systemic risks, biases, or catastrophic failures.

The global public good concept emphasizes how addressing transnational challenges can align with both national interests and global interests, potentially motivating governments to commit their own resources and collaborate on shared solutions that markets tend to under-provide—as in the case of climate change, where fossil fuels continue to dominate despite their negative externalities, and COVID-19, where critical medical supplies and equitable vaccine distribution were under-provided.[15]

[13] Zedillo, Ernesto, and Tidjane Thiam. "Meeting Global Challenges: International Cooperation in the National Interest." Meeting Global Challenges. Stockholm: International Task Force on Global Public Goods, 2006. https://ycsg.yale.edu/sites/default/files/files/Global-Public-Goods-expl.pdf also e.g. Stein, Felix, and Devi Sridhar. "Health as a 'Global Public Good': Creating a Market for Pandemic Risk." BMJ, August 31, 2017, https://doi.org/10.1136/bmj.j3397.

[14] Kaul, Inge, and Donald Blondin. "Global Public Goods and the United Nations." In *Global Governance and Development*, edited by José Antonio Ocampo. Initiative for Policy Dialogue Series. Oxford: Oxford University Press USA – OSO, 2016.

[15] Kaul, I. "Global Public Goods: Explaining Their Underprovision." Journal of International Economic Law 15, no. 3 (September 1, 2012): 729–50. https://doi.org/10.1093/jiel/jgs034.

This framework[16] may offer several distinct advantages for coordination and governance:

1. **Coordination Justification**: The framework provides economic and political justification for international coordination by demonstrating how addressing transnational challenges aligns with both national and global interests while simultaneously showing that uncoordinated action is unlikely to create desired outcomes.
2. **Investment Efforts**: The framing would highlight how market mechanisms or single state domestic governance measures alone could lead to the underprovision of the global public good, thus providing an argument for enhanced public investment.
3. **Institutional Architecture**: The framework suggests institutional arrangements and funding mechanisms for the provision of AI safety, drawing on precedents from other domains such as climate action and public health.
4. **Privileged Group Dynamics**: States may act as a "privileged group"—actors who possess sufficient resources, capabilities, and incentives to provide a public good and benefit, regardless of others' contributions or participation[17]— regarding certain global public goods, where pursuing their own interests might still bring global benefits.[18] This framing may influence how great powers approach AI safety: they might contribute to it as a global public good even while acting independently, for example by establishing and enforcing standards that primarily serve their interests while maintaining their technological advantages. At the same time, unilateral and uncoordinated decision-making by individual great powers may still induce further racing dynamics through competitive interactions and reduce the global goods produced by the privileged group's actions alone.

A crucial question emerges: which elements require protection or provision as global public goods? Just as clean air and climate stability represent essential global public goods for environmental governance, identifying core public goods in AI development provides a foundation for establishing effective governance mechanisms and coordination frameworks.

[16] Mazzucato, Mariana. "Governing the Economics of the Common Good: From Correcting Market Failures to Shaping Collective Goals." Journal of Economic Policy Reform 27, no. 1 (January 2, 2024): 1–24. https://doi.org/10.1080/17487870.2023.2280969.

[17] Olson, Mancur. The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, 1965. https://doi.org/10.4159/9780674041660.

[18] Ibid., 49–50.

Identifying Definitional Public Goods in the AI Realm

Established global public goods, such as clean air, disease control, and absence of transnational conflict, have emerged as clear bases for collective action. As AI systems become increasingly powerful and pervasive, must we identify analogous fundamental goods that require collective preservation? This question spans both AI's potential as a public good and the collective challenges posed by its development and deployment. Understanding these elements requires analyzing how different aspects of AI development and safety map onto existing global public goods frameworks.

Multiple Dimensions of Public Goods in Advanced AI

Fundamental human needs such as access to clean air and freedom from infectious disease, when framed as such in the political sphere, have animated collective responses to air pollution and pandemics.[19] However, multiple conferences or dialogues have identified analogous elements generating similar levels of concern in the context of advanced AI, across technical, social, and governance dimensions.

This section aims to examine the corresponding societal needs both for *access to* certain fundamental resources and circumstances, and also *freedom from* extremely adverse global conditions, in the specific context of advanced AI. Table 2 below provides a compact recap of analogous policy areas, distinguishing between core global public goods, corresponding societal needs, global public good characteristics, and complementary tools for advancing global public good provision. This will help us to apply the global public good framework to AI in the next section.

[19] See, for instance: Yin, Sonia, Xinyi Zhu, Yufei Cai, Jingqi Ju, and Shanxiao Liu. "The Effectiveness of Utilizing the Framing Effect to Motivate Climate Change Mitigation Effects on an Individual Level." Interdisciplinary Humanities and Communication Studies 1, no. 4 (January 3, 2024). https://doi.org/10.61173/cdk21542 and Pal, Leslie A. "Speaking Good to Power: Repositioning Global Policy Advice through Normative Framing." Policy and Society 42, no. 3 (October 12, 2023): 347–58. https://doi.org/10.1093/polsoc/puad012.

Table 2. Global Public Goods Comparison Framework

Issue Area	Core Universal Good	Access to / Provision of	Freedom from / Absence of	Facets of Non-Rivalry	Facets of Non-Excludability	Tools and Resources for GPG Provision
Climate	Stable environmental conditions and basic resource access	Clean air; Stable climate; Biodiversity; Environmental stability	Climate disasters; Environmental degradation; Resource scarcity; Extreme weather events; Climate-induced displacement	Benefits of climate stability shared by all; Mitigation efforts help everyone	Cannot exclude nations from climate effects; Atmosphere is inherently shared	Emissions reduction agreements
Global Peace	Global stability & security	International stability	War and conflict; Systemic violence	Peace benefits multiply with participation; Stability strengthens with broader adoption	Cannot exclude from stability benefits; Regional peace/conflict affects broader regions and alliances	International peacekeeping forces; Conflict mediation mechanisms; Arms control treaties; Early warning systems; Economic cooperation frameworks; Multi-stakeholder dialogue platforms; Joint security initiatives; Confidence-building measures; Dispute resolution mechanisms
Global Health (Pandemic prevention, specifically)	Conditions supporting human health	Medical treatments and vaccines; Basic health services; Healthcare infrastructure; Disease monitoring data	Pandemics; Health system collapse; Severe and widespread biological threats	Disease prevention helps all; Knowledge sharing enhances value; Research benefits multiply	Contagions cross borders regardless of national measures; Disease surveillance data benefits all regions; Health system failures affect neighboring countries; Pathogen evolution impacts global population	Disease prevention; Medical knowledge; Health systems; Research infrastructure; Early warning systems

How the Global Public Good Framing Has Been Applied to Artificial Intelligence in International Frameworks

The complexity of identifying fundamental public goods in AI development is reflected in how different international actors have approached this challenge. Recent international dialogues have produced several significant statements that attempt to map AI-related public goods, each highlighting different aspects and approaches to collective action (see Table 3). These have varied in focus, ranging from emphasizing safety protocols and verification mechanisms as potential non-rivalrous, non-excludable goods to the possible benefits of shared governance frameworks. The most prominent international dialogues and statements thus far have included the following:

- The **International Dialogues on AI Safety (IDAIS) Venice Statement**[20] approaches the question of fundamental public goods primarily through the **lens of technical safety measures**. It positions safety protocols and verification mechanisms themselves as non-rivalrous, non-excludable benefits, analogous to how clean air serves as a fundamental public good for climate stability. This framing emphasizes how safety measures, once developed, could theoretically benefit all without diminishing their value to any particular user.

- In contrast, the **Manhattan Declaration on Inclusive Global Scientific Understanding of AI**[21] takes a broader view, identifying scientific knowledge itself as the fundamental public good. This approach suggests that our collective understanding of AI capabilities, opportunities, and risks constitutes a shared resource that grows more valuable with broader participation, similar to how medical knowledge serves as a public good in global health frameworks.

- The **AI Safety as Global Public Goods Report**[22] adopts a more comprehensive perspective, framing governance capabilities themselves as fundamental public goods. This approach highlights how shared governance frameworks, like those for nuclear safety or aviation, can provide non-rivalrous benefits through cross-border policy learning and coordination. It particularly highlights the importance of balanced development in creating sustainable governance structures.

- **The UN High-Level Advisory Body on AI Report: Governing AI for Humanity**[23] recommendations take perhaps the broadest view, identifying multiple interconnected layers of public goods in the AI context. This framework encompasses not only technical

[20] International Dialogues on AI Safety. "IDAIS-Venice," September 5, 2024. https://idais.ai/dialogue/idais-venice/.

[21] "Mila's Yoshua Bengio, Alondra Nelson and Many Other AI Experts, Put Forward the Manhattan Declaration." Montreal Institute for Learning Algorithms, September 22, 2024. https://mila.quebec/en/news/milas-yoshua-bengio-alondra-nelson-and-many-other-ai-experts-put-forward-the-manhattan.

[22] Wang, Y., Jia, K., Zhao, J., Chen, L., Qin, C., Yuan, Y., Fu, H., Liang, X., et al. (2024). AI Safety as Global Public Goods Working Report. https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20241127/AI%20Governance%20as%20Global%20Public%20Commons.pdf.

[23] United Nations. Governing AI for Humanity: Final Report. New York, NY: United Nations, 2024. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

safety and scientific knowledge but also institutional capacity and development infrastructure as fundamental public goods requiring collective action.

See ***Appendix A*** for a more thorough comparison.

While all of the above statements and reports recognize the potential value of applying the global public good framework to the realm of AI, an in-depth analysis of these varying conceptualizations provides important insights into how domestic labs, state governments, and international institutions each uniquely understand the relationship between AI and global public goods. Thus, it is necessary to apply the framework developed in Table 2 to these existing international statements on AI and global public goods, as well as additional framings discussed during recent workshops with peer reviewers for this paper. In the resulting table below, we explore possible ways to conceptualize global public goods from advanced AI, their key characteristics, and supporting tools, resources, and processes.

When applying the global public goods framework to AI safety specifically, the distinction between an analytical framework and implementation tools becomes particularly salient. Understanding AI safety as a global public good reveals specific market and coordination failures that occur when individual actors underinvest in safety measures whose benefits extend beyond their immediate control. For instance, while a robust AI safety framework might generate non-rivalrous and non-excludable benefits globally, achieving this outcome depends on specific tools and resources—from technical standards to governance frameworks—that may themselves be excludable or rivalrous. This analysis helps explain why different framings of AI safety as a global public good might emphasize different sets of tools and resources for its provision.

Table 3. Alternative Framings of AI and Global Public Goods[(a)]

AI GPG Framing	Core Universal Good	Access to / Provision of	Freedom from / Absence of	Characteristics of Non-Rivalry	Characteristics of Non-Excludability	Tools for GPG Advancement
Technical Safety	Safe and controllable AI systems	Technical safety protocols; Verification mechanisms; Testing frameworks	Uncontrolled AI systems; Catastrophic failures; Systemic risks	Safety protocols benefit all users equally; Standards improve with wider adoption	Technical standards can be shared globally; Safety benefits extend across borders	Safety assessment frameworks; Verification protocols; Emergency response systems
Scientific Understanding	Shared knowledge of AI capabilities and risks	Research infrastructure; Scientific cooperation; Knowledge sharing platforms	Information asymmetries; Fragmented understanding; Isolated research	Scientific insights multiply through sharing; Research benefits from diverse inputs	Knowledge can be openly accessed; Scientific findings benefit all	Research collaboration networks; Open science platforms; Shared research infrastructure
Governance capabilities	Effective AI governance systems	Governance frameworks; Policy coordination; Stakeholder engagement	Governance failures; Regulatory gaps; Coordination failures	Governance knowledge benefits all parties; Best practices improve with sharing	Cross-border policy learning; Shared governance benefits	Policy coordination platforms; Multi-stakeholder frameworks; Governance standards
Development Infrastructure	Equitable AI development capacity	Technical infrastructure; Training resources; Development tools	Digital divides; Capability gaps; Resource inequities	Infrastructure benefits multiply with use; Knowledge sharing enhances value	Basic AI capabilities available to all; Development resources openly accessible	Capacity building networks; Resource sharing platforms; Development frameworks
Human Agency[(b)]	Human autonomy and dignity	Meaningful human choice and control; Decision-making agency	Automated oppression; Loss of human autonomy; Algorithmic discrimination	Protection of agency benefits all equally; Safeguards strengthen with broader adoption	Impact on human agency affects all; Cannot exclude from benefits of protected autonomy	Human oversight mechanisms; Agency protection frameworks; Rights preservation tools
Safety Knowledge Commons[(b)]	Collective AI safety expertise	Safety research; Best practices; Risk assessment tools	Knowledge gaps; Safety failures; Systemic risks	Safety knowledge grows with use; Benefits from diverse inputs	Safety insights benefit all; Cannot exclude from knowledge benefits	Knowledge sharing platforms; Collaborative research tools; Best practice repositories
Beneficial AI Systems[(b)]	AI systems serving the interests of humanity	Safe and beneficial AI applications; Public interest AI tools	Harmful AI applications; Misaligned systems; Negative externalities	Benefits from AI applications can be shared; Value grows with adoption	Basic AI benefits available to all; Cannot exclude from foundational benefits	Public interest AI development; Benefit sharing frameworks; Application standards

(a) Table notes: 1) This table synthesizes different conceptualizations of AI/AI safety as global public goods, drawing from international statements, workshop discussions, and academic analysis. 2) Each framing emphasizes different aspects of what constitutes the core universal good and how it exhibits public good characteristics. 3) The "Tools for GPG Advancement" column identifies specific mechanisms for operationalizing each framing. 4) Framings are not mutually exclusive; effective governance may require integrating multiple approaches.

(b) Additional angle first invoked in December 2024 paper workshop by workshop attendees.

Implications for Public Goods Analysis

These varying conceptualizations of global public goods suggest that rather than seeking an advanced AI-relevant singular analogue for messaging akin to other realms' global public goods like "clean air" or "disease prevention" that could motivate public action, multiple interconnected layers of public goods and tools are relevant to advanced AI:

1. **At the Technical and/or Built Level**: Basic safety and reliability of AI systems themselves represent fundamental public goods, analogous to clean air or stable climate. Open-source evaluation tools, testing frameworks, and safety protocols can be shared widely without losing value. However, unlike natural commons, these are human-created and require active maintenance. Additionally, technical measures for advancing the safety and reliability of AI systems are often closely intertwined with capability-enhancing technical tools and measures, potentially compromising the non-rivalry and non-excludability characteristics of these specific measures in contexts of interstate competition and comparative anxieties held at the state level around advanced technology-based capabilities. However, this does not necessarily compromise the non-excludability and non-rivalry characteristics of the core global public good. This tension between open and proprietary approaches to safety reflects broader challenges in balancing collective benefit with innovation incentives.
2. **At the Knowledge Level**: Scientific understanding and governance capabilities may constitute another layer of public goods, or they may also fall into the category of tools and resources to advance the provision of the core public good of AI safety, similarly to medical knowledge in public health. These grow more valuable with broader participation but face challenges of access and equity; the expertise needed to develop and apply this knowledge remains concentrated in specific institutions and regions.
3. **At the Institutional Level**: Governance frameworks and capacity development infrastructure form a third layer, comparable to international frameworks aimed at promoting peace or preventing disease. In AI safety, best practices for safety governance and incident response protocols demonstrate strong public good characteristics. These organizational frameworks can be adapted and implemented across different contexts without diminishing their effectiveness. However, the specialized expertise and institutional capacity needed to implement these practices effectively may be rivalrous and temporarily excludable.

This layered understanding helps explain why some aspects of AI development exhibit clear non-rivalry and non-excludability while others remain tied to excludable capabilities or rivalrous resources. Specifically, countries might want to keep these tools private to maintain technological advantages over other nations, even though sharing them would help make AI safer for everyone. It also suggests why different international statements emphasize different aspects of the public goods framework. These multiple dimensions suggest that framing AI safety as a global public good requires careful consideration of how different aspects align with traditional public goods characteristics. While some elements of AI safety exhibit clear non-rivalry and non-excludability, other framings or tools to achieve them may be inherently linked to excludable capabilities or rivalrous resources. Understanding these nuances is crucial for developing effective governance frameworks that can address both immediate safety challenges and longer-term societal implications.

Several promising framings emerge from these insights:

1. **Human Agency as Universal Good**: Human agency and autonomy represent basic conditions for human flourishing in an AI-enabled world. The preservation of meaningful human choice and control could serve as a universal good that, once compromised, affects all of humanity regardless of individual circumstances or contributions.
2. **Safety Knowledge as Shared Resource**: Similar to how medical knowledge functions as a public good in global health, collective understanding of AI safety could represent a non-rivalrous, non-excludable resource that grows more valuable with broader participation and sharing and supports the advancement of a core global public good that focuses on human well-being or agency.
3. **Governance Capability as Common Infrastructure**: Like international governance structures for aviation safety or nuclear security, shared governance capabilities for AI could provide universal benefits while improving with broader adoption and implementation.

Table 3 above can thus help identify specific borderless risks that all humanity faces, such as:

- Loss of meaningful human agency and autonomy;
- Systemic failures in AI-dependent infrastructure;
- Catastrophic incidents from uncontrolled AI systems;
- and governance failures in managing powerful AI capabilities.

Furthermore, the development of shared safety frameworks and assessment tools could provide universal benefits—analogous to how clean air is an identifiable global public good in the climate change mitigation case—by:

- Protecting fundamental human capabilities and rights;
- Ensuring reliable and controllable AI systems;
- Enabling effective governance and coordination;
- and supporting equitable development and deployment.

While no single framing perfectly captures all aspects of AI safety as a global public good, combining multiple perspectives can help establish clear focal points for collective action. This analysis is not meant to express a definitive opinion about which framing is best, or even to determine a number of framings that should be considered, but rather to spark discussion and prompt further research.

Applying the “Global Public Good” Framing to AI Safety

Certain domains of knowledge, capabilities, and resources related to AI safety exhibit characteristics of non-rivalry and non-excludability, which suggests that AI safety might be framed as a global public good. Building on the theoretical framework established earlier, this section evaluates how the global public goods concept specifically applies to AI safety—the measures and mechanisms aimed at preventing accidents, misuse, or other harmful consequences from AI systems. While various stakeholders interpret “AI safety” differently, this analysis focuses on concrete measures that ensure that AI systems operate safely and reliably, from preventing immediate harms (such as systemic biases and privacy violations) to addressing longer-term risks to human agency and societal stability.

While some critical resources needed for AI safety work (such as compute infrastructure, research funding, and specialized talent) are both rivalrous and excludable, the safety and capabilities knowledge produced through safety research often display clear public good characteristics. The non-excludability of AI safety manifests primarily through the inherently global nature of both AI capabilities and their associated risks. Safety measures, once developed and implemented, naturally extend their protective benefits beyond the immediate jurisdiction or entity that created them. For instance, advances in interpretability techniques or robustness measures typically generate knowledge that, while it may be temporarily restricted through intellectual property protections, ultimately influences practices across the AI safety field (in which open-access preprints and open-source innovations have historically been norms). This dynamic parallels other global public goods such as climate stability, where the benefits of protective measures cannot be meaningfully restricted to specific nations or regions.

While the global public goods framework offers promising theoretical foundations for addressing collective challenges in AI development and safety, its practical application demands careful consideration of both benefits and limitations. The framework’s potential to facilitate international coordination and justify collective action must be weighed against complex implementation challenges and political realities. This section examines how the global public goods framing operates across different levels of governance, explores key tensions in responsibility allocation, and analyzes practical political considerations that could impact its effectiveness as a governance tool. Understanding these dynamics is crucial for developing viable approaches to ensuring AI safety while navigating the complex landscape of international cooperation and competing national interests.

Regional, National and Global Public Goods in AI Safety Governance

Historical examples from nuclear safety, aviation, and food safety demonstrate how public goods often emerge first at regional or national levels before expanding to an international scope. Nuclear safety, for instance, developed primarily through national regulatory frameworks, with events such as the Three Mile Island incident driving domestic safety improvements in the US before eventually contributing to international standards.[24] Similarly, aviation safety evolved from local and national concerns about specific carriers into robust international frameworks for aircraft certification and airline operation.[25] Food safety followed a comparable pattern, with local incidents such as contamination outbreaks driving regional responses that gradually contributed to international standards.[26]

These historical patterns suggest important insights for AI safety governance. While advanced AI systems may ultimately pose rapid, borderless risks, many immediate safety challenges manifest first at local or regional levels, such as infrastructure reliability or system testing protocols. Measures to address these localized risks often contribute to mitigating broader global risks, suggesting a potential pathway where national and regional safety frameworks serve as building blocks for international governance. This graduated approach aligns with historical patterns in technological governance, where public goods frameworks proved particularly effective at national and regional levels before scaling to global coordination.

Moreover, this "local first" approach means that policymakers can leverage existing governance capabilities while avoiding the dilution of responsibility that can occur when issues are immediately framed as global challenges. The graduated approach to public goods provision in AI safety could help bridge the gap between nations with significant AI development capabilities and those primarily concerned with managing AI's impacts, while maintaining clear lines of responsibility for safety measures at each level of governance.

Different types of global goods have given rise to different potential pathways for addressing challenges around incentives, though not all of these approaches work for every kind of global public good. This breakdown sheds light on how different components of the overall AI safety field could align with each approach as part of a whole strategy:

[24] Walker, J. Samuel. Three Mile Island: A Nuclear Crisis in Historical Perspective. Berkeley: University of California Press, 2004.

[25] Priest, W. Curtiss. Risks, Concerns, and Social Legislation: Forces That Led to Laws on Health, Safety, and the Environment. 1st ed. Routledge, 2019. https://doi.org/10.4324/9780429304897.

[26] Priest, W. Curtiss. Risks, Concerns, and Social Legislation: Forces That Led to Laws on Health, Safety, and the Environment. 1st ed. Routledge, 2019. https://doi.org/10.4324/9780429304897.

1. **Aggregate Efforts** require coordinated contributions across stakeholders to get at a sufficiently broad and deep target only made possible through cumulative work. In AI safety, this could manifest through collaborative development of safety assessment standards, similar to how aerosol/anti-chloro-fluoro-carbon regulation required comprehensive participation from producing and consuming nations. For this approach to succeed, policymakers must establish clear standards and ensure broad participation in their development and implementation, especially to avoid "regulatory flight": the phenomenon of AI developers moving activities to areas with less burdensome rules.[27]

2. **Weakest Link**: This framework focuses on addressing vulnerabilities in the global system's weakest points. For AI safety, this suggests prioritizing basic safety capabilities and monitoring systems in all jurisdictions where AI development or deployment occurs. The approach recognizes that unsafe AI development in any location could generate risks for all, similar to how disease control requires comprehensive global surveillance since diseases arise in one location but can cross borders.

3. **Single Best-Shot**: This model focuses on breakthrough solutions that, once developed, can benefit all stakeholders. In AI safety, this might apply to fundamental advances in areas such as formal verification methods or robustness techniques that, once discovered, could be widely implemented. This approach suggests that we should concentrate resources on key technical challenges while ensuring that the solutions can be distributed globally.

The selection and combination of approaches should reflect both technical realities and political feasibility. This graduated approach could help build momentum while maintaining clear lines of responsibility and accountability.

Political Realities, Power Dynamics, and Implementation Pathways

The framing of AI safety as a global public good requires careful examination of practical political implications, particularly given current geopolitical dynamics and historical experiences with similar frameworks. A central tension emerges between concentrated and diffused responsibility in the global public goods framing. In current discourse, AI safety is often reasonably framed as the responsibility of advanced AI organizations and leading states, which suggests clear moral arguments for those actors to take direct action. The global public goods framing, on the other hand, potentially risks diluting this focused accountability. The incentive challenge manifests in the misalignment between individual and collective interests in safety investment. While all stakeholders benefit from robust AI safety measures, individual actors may lack sufficient motivation to invest adequately, particularly when safety advances might be shared globally while development costs remain local. This parallels challenges observed in climate

[27] Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, et al. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." arXiv, November 7, 2023. https://doi.org/10.48550/arXiv.2307.03718.

change mitigation, where individual nations may hesitate to bear costs for emissions reductions that benefit all countries. This section also explores the entanglement of AI safety work and capabilities enhancements.

National Interests and Inequalities

It has been challenging to apply global public goods frameworks when faced with entrenched national interests, unequal power dynamics, and substantial financial costs.[28] States face significant hurdles coordinating joint action on AI safety due to differing national interests, varying technical capabilities, and a lack of established international governance mechanisms for emerging technologies. Individual states may also be reluctant to heavily invest in AI safety measures when they fear their strategic competitors might instead direct those resources toward developing offensive AI capabilities or economic advantages, creating a prisoner's dilemma where the collectively optimal outcome of prioritizing safety is undermined by competitive pressures. Contemporary resistance to global public goods frameworks often centers on sovereignty concerns, with states demonstrating particular reluctance to accept supranational governance mechanisms.[29] This resistance manifests in complex debates about resource allocation, including questions of funding distribution and concerns about free-riding behavior. Implementation challenges further complicate the picture, particularly regarding verification mechanisms and compliance monitoring.

Furthermore, postcolonial critiques of climate change governance have demonstrated how seemingly neutral international frameworks can function as sophisticated mechanisms of economic and normative control.[30] The terminology of "global public goods" itself often reinforces existing geopolitical hierarchies despite intentions of promoting collective responsibility by, for instance, being invoked within international treaty mechanisms that enforce uniform responsibility for an unequally-created problem, or enacting top-down policies dictated by the most powerful actors globally.[31] Additional critiques toward the global public good

[28] Stückelberger, Christopher. "Post-Corona World: Balancing International Cooperation and National Sovereignty." Journal of Law and Administration 16, no. 2 (June 26, 2020): 10–17. https://doi.org/10.24833/2073-8420-2020-2-55-10-17.

[29] Zedillo, Ernesto, and Tidjane Thiam. "Meeting Global Challenges: International Cooperation in the National Interest." Meeting Global Challenges. Stockholm: International Task Force on Global Public Goods, 2006. https://ycsg.yale.edu/sites/default/files/files/Global-Public-Goods-expl.pdf.

[30] This justice-based approach to analyzing the differential burdens of global public good provision (in the climate change case) is summarized in Climate Change and Society: Sociological Perspectives edited by Riley E. Dunlap, Robert J. Brulle: "If historical responsibility is taken into account, Global North nations have consumed more than three times their share of the atmosphere (in terms of the amount of emissions that we can safely put into the atmosphere) while the poorest 10 percent of the world's population has contributed less than 1 percent of CO2 emissions. [...] [However,] the Global South and people of color, Indigenous communities, the poor, and women and children in all nations are precisely the populations that bear the brunt of climate disruption in terms of its ecological, economic, and health burdens. [...] Many governments of the Global South feel strongly that they have paid a heavier price for climate change, or are likely to pay as impacts grow worse, while receiving very few of the benefits." Dunlap, Riley E., Robert J. Brulle, and American Sociological Association, eds. Climate Change and Society: Sociological Perspectives. New York, NY: Oxford University Press, 2015 (pp.127–128).

[31] See, for example, this longform dissection of the Paris Agreement process, and the ways in which powerful states

framing include how the focus of the core concept is aimed at addressing market failures only where private sector incentives are insufficient, which would restrict the role of the state to that of a market facilitator rather than a provider and shifting attention away from the structural inequities and access barriers core to the neoliberal status quo.[32]

This dynamic is particularly relevant for AI safety, since AI development capabilities are highly concentrated, and framing AI safety in market failure terms rather than using, for instance, a more rights-based lens. The asymmetry between the small number of nations driving AI development and the global scope of potential impacts creates complex governance challenges. While developing nations need to be included in safety governance structures, they may express legitimate concerns about safety requirements potentially limiting or slowing down access to AI-fueled or AI economy-fueled development opportunities.[33] The perception that safety measures might slow down technological progress presents a particular challenge, especially for developing nations, which parallels concerns raised during climate change negotiations, where developing nations feared that emissions restrictions could hamper industrialization and economic growth, leading to the principle of "common but differentiated responsibilities" in the UN Framework Convention on Climate Change. While some critics have pointed out that distributional concerns might be offset by the potential transformative impact of safe AI for all economies, developing nations likely cannot justify ignoring distributional concerns with the promise of future AI-driven abundance, since without immediate representation in AI development and safety frameworks, they risk being locked out of both near-term benefits and long-term governance decisions, potentially deepening existing global inequalities.

While most risks may originate from a handful of countries, the consequences of those risks and the need for safety measures affect the entire global community. This disparity in capability and impact creates complex dynamics for international cooperation and responsibility allocation, highlighting the importance of incorporating capacity building and knowledge transfer mechanisms into any global governance framework. As a result, at international fora and toward domestic political leadership, civil society's framing of AI safety as a global public good requires careful attention to communication and trust-building across stakeholder groups.

When the public-private nexus is taken into consideration, the global public goods framing might also inadvertently discourage private sector investment in safety research, as companies

pushed to reduce legally binding forms of action and mitigation, as well as internal transparency measures, that were of particular importance to smaller states more immediately at risk in the near-term from climate change-derived catastrophes. Dimitrov, Radoslav S. "The Paris Agreement on Climate Change: Behind Closed Doors." Global Environmental Politics 16, no. 3 (August 2016): 1–11. https://doi.org/10.1162/GLEP_a_00361.

[32] For more of this critique, see: Saksena, Nivedita. "Global Justice and the COVID-19 Vaccine: Limitations of the Public Goods Framework." Global Public Health 16, no. 8–9 (September 2, 2021): 1512–21. https://doi.org/10.1080/17441692.2021.1906926.

[33] "This global disparity in AI rule-setting means that the technology's path will trace the national, commercial, and social interests of wealthy nations, at times to the detriment of societies with less power and fewer resources in the global South. Without a greater say and more AI policymaking capacity, these populations are more likely to be exposed to AI risks and deprived of AI benefits." From: LaForge, Gordon. "The Dangers of Imposing Global North Approaches to AI Governance on the Global South." Tech Policy Press, September 5, 2024. https://techpolicy.press/the-dangers-of-imposing-global-north-approaches-to-ai-governance-on-the-global-south.

would be expected to share advances while bearing development costs. This could exacerbate free-rider problems and make it harder to justify concentrated safety efforts at leading research institutions. Thus, the framework needs to address how to fairly distribute the costs and benefits of safety research across the global AI development ecosystem.

Overall, given these political realities, successful instantiation of AI safety as a globally-recognized public good will likely require an approach that builds on disparate aspects of AI safety work at the local, regional, and global levels, while developing scalable governance mechanisms held in common across those levels. It will also require actors to preserve a sense of urgency as increasingly powerful models and capabilities emerge, given that retrofitting safety measures and coordinating responses after deployment will likely be far more difficult. This approach should enable capacity building across regions, allowing us to share technical knowledge and best practices while creating inclusive governance structures. Innovation in institutional design will be crucial. It is particularly important to develop new coordination practices and establish clear accountability mechanisms that respect varying national priorities and capabilities.

Safety and Capabilities Work

A fundamental challenge in framing AI safety as a global public good also lies in the complex relationship between safety advances and capability developments in AI systems. Safety measures often require sophisticated technical capabilities to implement effectively, creating what we term the “safety-capability paradox.” This interconnection makes it difficult to advance safety measures without simultaneously enabling capability improvements that may carry their own risks. For instance, improvements in model interpretability might simultaneously enhance safety oversight and enable more sophisticated (and more risky) AI applications. This dynamic complicates efforts to share safety advances globally while managing capability proliferation concerns. As a result, the risk of “safety-washing”—where actors use safety-based language to describe advances in capabilities, leading to confusion and misperceptions of the steps actually being taken—presents another political challenge that could undermine the credibility of global public goods approaches.[34]

The challenge of distinguishing safety-specific work from capability-enhancing research further complicates this picture. Some aspects of AI safety work can be clearly separated from capability advancement:

- Monitoring and oversight tools that detect harmful content or discriminatory patterns
- Testing protocols and evaluation frameworks for assessing safety properties
- Organizational safety measures such as incident response protocols and documentation standards

[34] Ren, Richard, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, et al. “Safety-washing: Do AI Safety Benchmarks Actually Measure Safety Progress?” arXiv, December 27, 2024. https://doi.org/10.48550/arXiv.2407.21792.

- Transparency tools for auditing and explaining AI behavior
- Basic safety architecture elements such as shutdown mechanisms and access controls

However, other crucial areas of safety research remain tightly coupled with capability advancement:

- Technical alignment research requiring advanced AI understanding
- Evaluation frameworks that require sophisticated AI reasoning
- Reliability improvements that enhance system performance
- Scalable oversight mechanisms that may advance general capabilities

This entanglement creates practical challenges for implementing global public goods frameworks, as sharing safety advances might inadvertently accelerate capability development in ways that raise security concerns, especially if risks from advanced AI are deemed by a given state to be less important than the risks of advanced AI capabilities in the hands of that state's competitors.

Conclusions on the Politics of AI Safety as a Global Public Good

In short, major AI powers may resist sharing safety advances that they perceive as conveying strategic advantages, while developing nations may view safety requirements as potential barriers to their own AI development aspirations, where speed of implementation or deployment could provide competitive economic benefits over whatever is comparatively gained from voluntarily focusing on more robust safety measures. The former concern can sometimes be offset by advanced AI-developing countries who are not involved in any rivalries and who can work as safety entrepreneurs, though currently the states with most advanced AI capabilities are also involved in strategic competition against one another.

The practical political landscape suggests several alternative or complementary ways to frame AI safety. A national security framing of the importance of AI safety can increase attention and resource allocation but may narrow the scope of the risks that are considered and complicate international collaboration. Regional cooperation offers advantages, since different regions can build on their existing governance structures which allows for a graduated approach to international coordination. Public education and engagement represent another crucial dimension, paralleling successful aspects of climate change advocacy. Building broader understanding of AI safety can create pressure for political action and help establish the legitimacy of global governance frameworks. However, this approach requires careful balance to avoid oversimplifying complex technical issues or creating unwarranted panic.

Overall, these challenges suggest the need for nuanced approaches that:

1. Carefully delineate aspects of safety work—such as interpretability tools, robustness testing frameworks, and oversight mechanisms—that can be credibly separated from capability advancement
2. Develop mechanisms for sharing safety advances that do not compromise security interests—i.e. potentially through trusted intermediaries or selective disclosure protocols that protect sensitive technical details while still allowing broad adoption of key safety principles and methods
3. Create incentive structures that maintain private sector investment while promoting collective benefit
4. Address equity concerns while maintaining clear lines of responsibility

While these challenges are significant, they are not necessarily insurmountable. Solutions ought to be matched to the nature of the problem: if safety components cannot be accomplished without international coordination, time and resources would be wasted focusing primarily on the actions of specific individual firms, for instance; or, if a safety component can be effectively implemented by individual organizations, it could also waste time and resources to create international agreements about it. Success likely requires carefully structured approaches that acknowledge and address these limitations while leveraging the framework's benefits.

Research Agenda: Global Public Goods and Advanced AI

This paper has examined the concept of AI safety as a global public good, analyzing both the potential and limitations of this framing. Through our analysis of recent international statements, theoretical frameworks, and emerging governance challenges, several key themes have emerged. First, while the global public good framing offers valuable perspectives for addressing collective action problems in AI development, it also faces significant practical and theoretical challenges in implementation. Second, the evolving nature of AI capabilities means that our understanding of what constitutes "safety" as a public good must remain dynamic and responsive. Third, existing international frameworks and institutions may need substantial adaptation—or entirely new mechanisms may need to be developed—to effectively govern AI safety as a global public good.

Rather than advocating for specific policy measures at this early stage, this analysis suggests the need for a robust research agenda to inform future governance efforts. This agenda must address both immediate practical challenges and longer-term theoretical questions about the nature of public goods in an AI-enabled world. Below, we outline key areas that require further research and discussion across policy and AI communities.

Areas for further research:

Core Universal Needs

- How should we identify and prioritize potential global risks from advanced AI that require collective action?
- Which specific aspects of human agency and autonomy are most threatened by advanced AI systems, and how do these threats manifest across different development contexts?
- What minimum safety requirements must be universally guaranteed to prevent catastrophic AI accidents or misuse?
- How do these requirements vary across different cultural, economic, and political contexts?
- What critical freedoms must be protected? What security provisions are essential? What resources do global populations need access to?
- What are the most productive and necessary focal points of the global public good framing, and what tools, resources, and policy efforts are needed to advance their provision?

Elements of AI Safety and Global Public Good Framing

- In what ways do AI robustness measures and testing frameworks constitute public goods? Are there occasions where benefits are excludable (through patents and trade secrets), yet their failures could have non-excludable negative consequences?
- How should AI transparency and monitoring systems be classified in terms of public goods, and do these function differently at national versus global levels, particularly given varying international standards and capabilities?
- To what extent can AI safety standards and governance frameworks be considered public goods? How do their implementations vary between being a national public good versus requiring global coordination for effectiveness?

Governance Levels and Distribution

- How can governance frameworks balance the concentrated capability and responsibility of leading AI developers with the need for inclusive international participation?
- What mechanisms could help bridge the stark disparity between states with primary involvement in developing advanced AI versus states who mainly consume advanced AI created in other nations while maintaining robust safety standards?
- How should safety protocols be adapted across different development contexts without compromising their effectiveness?

Governance Mechanisms

- What new governance mechanisms could effectively manage the "safety-capability paradox", where safety measures generate new intertwined capabilities insights or may require advanced capabilities to implement?

- How can verification mechanisms be designed to enable the sharing of safety advances while protecting legitimate security interests?

- How can meaningful public debate and input be facilitated, while acknowledging power asymmetries between states with primary involvement in developing advanced AI versus states who mainly consume advanced AI created in other nations?

- How can governance mechanisms avoid perpetuating existing inequities, enable meaningful participation from all affected communities, balance competing interests and needs, and protect against exploitation or marginalization?

Implementation and Actor Response

- How do different stakeholders interpret and respond to the "global public good" framing of AI safety across various cultural and political contexts?

- What factors influence organizations' willingness to participate in international coordination mechanisms when AI safety is framed as a global public good?

- To what extent has the framing of global public goods driven collective action?

- What impact would framing AI safety as a "global public good" have on different stakeholders' willingness to participate in international coordination mechanisms? How does this vary across government agencies, international organizations, industry leaders, and civil society groups? What factors influence this receptiveness?

- How do different cultural and political contexts affect how different stakeholders interpret and respond to the "global public good" framing, specifically as applied to AI safety?

Measurement and Metrics

- How might we develop reliable metrics and auditing mechanisms to distinguish genuine safety improvements from superficial "safety-washing" or capability-oriented work?

- What metrics could help assess whether safety measures are effectively functioning as public goods?

- How can we evaluate the distribution and accessibility of safety benefits across different regions and stakeholders?

Technical and Economic Considerations

- What technical mechanisms could help advance critical safety advances while managing the risks of capability proliferation?
- How can economic incentives be structured to encourage private sector investment in safety while promoting collective benefit?
- What funding mechanisms could address the free-rider problem in safety research while ensuring equitable access to safety measures?
- How can we effectively evaluate distributional impacts of governance approaches, effectiveness of equity measures, and power dynamics in decision-making and accountability?

Closing Reflections

The challenge of ensuring AI safety represents one of the most significant governance challenges of our time. While this paper has identified key areas for research and discussion, success will require sustained engagement across disciplines, sectors, and national boundaries. It will demand new thinking about governance, new forms of international cooperation, and new approaches to balancing competing interests and needs. Moreover, the accelerated pace of AI development has often dictated the implementation of safety measures before full empirical validation, while also demanding that governance and safety are done right from the outset, since the window for corrective action may diminish or disappear altogether once advanced systems are deployed.

The research agenda outlined above is not exhaustive but rather represents priority areas where focused investigation could help inform policy development and governance design. As AI capabilities continue to advance, the urgency of addressing these questions grows. Progress will require both theoretical insight and practical experimentation, combined with ongoing dialogue between technical experts, policy makers, and affected communities worldwide.

Ultimately, the goal must be to ensure that advanced AI systems contribute to human flourishing while protecting against potential harms. This requires treating AI safety not merely as a technical challenge but as a fundamental public good requiring coordinated global action. Success in this endeavor could help establish precedents and mechanisms for addressing other global challenges, while failure could have profound consequences for human welfare and development.

Appendix A: Statements on AI Safety as a Global Public Good

This appendix examines key international statements that have addressed AI governance through global public goods frameworks. While these statements share some common elements in their approach to international cooperation, they differ significantly in their primary focus, conceptual frameworks, and proposed implementation mechanisms. Understanding these distinctions is crucial for developing effective international governance approaches.

Recent International Developments

The concept of AI safety as a global public good has gained significant traction in international fora:

- **International Dialogues on AI Safety**: A series of meetings bringing together top AI scientists from China and the West, including Turing Award winners Yoshua Bengio, Andrew Yao, and (now also Nobel Laureate) Geoffrey Hinton. Their most recent meeting in Venice (September 2024) produced a consensus statement stressing the urgent need for global cooperation on AI safety and included specific recommendations, such as establishing emergency preparedness agreements and institutions, developing a safety assurance framework, and promoting independent global AI safety and verification research through establishing funds. The statement included the conclusion that "**The global nature of these risks from AI makes it necessary to recognize AI safety as a**

global public good, and work towards global governance of these risks. Collectively, we must prepare to avert the attendant catastrophic risks that could arrive at any time."[35]

- **Manhattan Declaration on Inclusive Global Scientific Understanding of AI**: Signed at the UN General Assembly in September 2024, this declaration by 21 influential AI researchers and policy professionals aims to promote AI as a "global public good" and encourages inclusive, global approaches to understanding AI's capabilities, opportunities, and risks. The declaration stated, "We reaffirm our commitment to developing AI systems that are beneficial to humanity and acknowledge their pivotal role in attaining the global Sustainable Development Goals, such as improved health and education. We emphasize that AI systems' whole life cycle, including design, development, and deployment, must be aligned with core principles, safeguarding human rights, privacy, fairness, and dignity for all."[36]

- **UN High-Level Advisory Body on AI Report: Governing AI for Humanity**: Released recommendations for promoting responsible and safe AI governance globally, including an international scientific panel on AI, AI standards exchanges, a capacity development network, and a global fund for AI to mitigate the widening "AI divide." It also stated that, "Pooling scientific knowledge is most efficient at the global level, **enabling joint investment in a global public good**, and public interest collaboration across otherwise fragmented and duplicative efforts."[37]

- **AI Safety As Global Public Goods Report (EN translation)**: The "AI Safety as a Global Public Goods" Chinese report was released at the Shanghai World AI Conference in July 2024. It acknowledges the positive and significant role of multilateral and multi-stakeholder actions in advancing AI safety, while also recognizing potential areas for improvement.[38]

Comparison of Conceptual Structures

Each statement conceptualizes the relationship between AI and global public goods differently:

1. International Dialogues on AI Safety Venice Statement
    - Focuses specifically on safety of advanced AI systems
    - Defines safety primarily through technical measures and verification

[35] International Dialogues on AI Safety. "IDAIS-Venice," September 5, 2024. https://idais.ai/dialogue/idais-venice/.

[36] "Mila's Yoshua Bengio, Alondra Nelson and Many Other AI Experts, Put Forward the Manhattan Declaration." Montreal Institute for Learning Algorithms, September 22, 2024. https://mila.quebec/en/news/milas-yoshua-bengio-alondra-nelson-and-many-other-ai-experts-put-forward-the-manhattan.

[37] United Nations. Governing AI for Humanity: Final Report. New York, NY: United Nations, 2024. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

[38] Wang, Y., Jia, K., Zhao, J., Chen, L., Qin, C., Yuan, Y., Fu, H., Liang, X., et al. (2024). AI Safety as Global Public Goods Working Report. https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20241127/AI%20Governance%20as%20Global%20Public%20Commons.pdf.

   - Emphasizes catastrophic risk prevention
   - Frames safety protocols themselves as global public goods

2. Manhattan Declaration on Inclusive Global Scientific Understanding of AI
   - Takes a broader view of scientific understanding
   - Emphasizes inclusive participation in knowledge creation
   - Links scientific cooperation to governance outcomes
   - Frames scientific knowledge itself as a primary public good

3. AI Safety As Global Public Goods Report
   - Adopts a comprehensive governance perspective
   - Includes safety alongside reliability and fairness
   - Emphasizes practical implementation
   - Frames governance capabilities as shared resources

4. UN High-Level Advisory Body on AI Report: Governing AI for Humanity
   - Takes a holistic approach to development and governance
   - Emphasizes capacity building and inclusive participation
   - Addresses multiple dimensions of AI development
   - Frames various aspects as public goods, from knowledge to infrastructure

Overview and Conceptual Structures

Table 4. Primary Characteristics of International Statements

Statement	Date	Primary Authors	Institutional Context	Focus Area	Global Public Good Framing
International Dialogues on AI Safety Venice Statement	Sept 2024	Leading AI scientists including Turing Award winners	Scientific consensus statement	Technical safety and catastrophic risk prevention	Safety measures and verification protocols as non-rivalrous, non-excludable benefits
Manhattan Declaration on Inclusive Global Scientific Understanding of AI	Sept 2024	Diverse group of AI scientists and policy researchers	Scientific-policy consensus statement	Inclusive scientific understanding of AI capabilities	Scientific knowledge and cooperation as shared global resources
AI Safety As Global Public Goods Report	July 2024	Government and academic institutions	Policy analysis document	Comprehensive governance including reliability and fairness	Knowledge sharing and governance capabilities as public goods
UN High-Level Advisory Body on AI Report: Governing AI for Humanity	Sept 2024	Multi-stakeholder expert group	International institutional framework	Inclusive development and balanced governance	Multiple dimensions including knowledge, standards, and capacity as public goods

Table 5. Conceptual Approaches to Global Public Goods

Statement	Non-rivalry Emphasis	Non-exclusivity Emphasis	Implementation Focus	Development Context
International Dialogues on AI Safety Venice	Safety protocols benefit all users equally	Technical standards available globally	Verification and emergency response	Focus on advanced AI systems and nations
Manhattan Declaration on Inclusive Global Scientific Understanding of AI	Scientific insights multiply with sharing	Open participation in research	Knowledge sharing and cooperation	Emphasis on inclusive participation
AI Safety As Global Public Goods Report	Governance knowledge benefits all parties	Cross-border policy learning	Multi-stakeholder coordination	Balanced development approach
UN High-Level Advisory Body on AI Report: Governing AI for Humanity	Multiple benefits from shared frameworks	Universal access to governance tools	Institutional capacity building	Strong focus on developing nations

Divergent Definitions and Emphases

The statements differ notably in how they define key concepts:

1. AI Safety and Risk
   - International Dialogues on AI Safety Venice: Technical safety measures and catastrophic risk prevention
   - Manhattan Declaration on Inclusive Global Scientific Understanding of AI: Broader conception including societal implications
   - AI Safety As Global Public Goods Report: Component of comprehensive governance
   - UN High-Level Advisory Body on AI Report: Governing AI for Humanity: One aspect of balanced development
2. Global Public Goods Relating to AI
   - International Dialogues on AI Safety Venice: Primarily technical protocols and standards
   - Manhattan Declaration on Inclusive Global Scientific Understanding of AI: Scientific knowledge and research cooperation
   - AI Safety As Global Public Goods Report: Governance capabilities and practices
   - UN High-Level Advisory Body on AI Report: Governing AI for Humanity: Multi-dimensional, including capacity and infrastructure
3. International Cooperation
   - International Dialogues on AI Safety Venice: Technical coordination and verification
   - Manhattan Declaration on Inclusive Global Scientific Understanding of AI: Scientific collaboration and knowledge sharing
   - AI Safety As Global Public Goods Report: Multi-stakeholder governance coordination
   - UN High-Level Advisory Body on AI Report: Governing AI for Humanity: Inclusive development and capacity building

Partner Organizations

Oxford Martin AI Governance Initiative

The AI Governance Initiative is co-led by Robert Trager, a social scientist specialising in international relations and frontier AI regulation, and Michael Osborne, a specialist in machine learning. Housed in the Martin School of the University of Oxford, AIGI is one of the few centres in the world focused on the governance of AI from both technical and policy perspectives. The initiative aims to anticipate and mitigate lasting risks from AI through (1) impactful research that is rigorously grounded in the social and computational sciences, (2) decision-maker education campaigns, and (3) training the next generations of technology governance leaders.

Concordia AI

AI is likely the most transformative technology that has ever been invented. Controlling and steering increasingly advanced AI systems is a critical challenge for our time.

Concordia AI is a social enterprise with offices in Beijing and Singapore focused on AI safety and governance. We aim to ensure that AI is developed and deployed in a way that is safe and aligned with global interests. We provide expert advice on AI safety and governance, support AI safety communities in China, and promote international cooperation on AI safety and governance.

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 170 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Technology and International Affairs Program

The Technology and International Affairs Program develops insights to address the governance challenges and large-scale risks of new technologies. Our experts identify actionable best practices and incentives for industry and government leaders on artificial intelligence, cyber threats, cloud security, countering influence operations, reducing the risk of biotechnologies, and ensuring global digital inclusion.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

安远 AI
CONCORDIA AI

AI GOVERNANCE
INITIATIVE

OXFORD
MARTIN
SCHOOL

UNIVERSITY OF
OXFORD