Given the Absence of Hard Law, the Roles for Soft Law Functions in the International Governance of AI

Wendell Wallach, Anka Reuel, and Anja Kaspersen

Abstract:

The advent of foundation models has alerted diplomats, legislators, and citizens around the world to the need for AI governance that amplifies benefits, while minimizing risks and undesired societal impacts. The prospects that AI systems might be abused, misused or unintentionally undermine international stability, equity, and human rights demands a high degree of cooperation, oversight, and regulation. However, governments are not acting quickly enough on putting in place an international hard law regime with enforcement authority. In the absence of such a regime, soft laws become a lever to help shape the trajectory of AI development and encourage international cooperation around its normative and technical governance. In this paper, we give an overview of key soft law functions in the context of international AI governance and mechanisms to fulfill them. We further propose the establishment of a Global AI Observatory in line with Mulgan et al. (2023) to fulfill functions that have not been (sufficiently) picked up by, or go beyond the mandate of, existing institutions.

Introduction

Recent calls for the international governance of AI[1] seldom include any specifics. A few governance initiatives are under development that have international ramifications, but they are driven by small groups of states. Those with experience in international policy recognize the serious difficulties that would be involved in putting in place new and inclusive international governance mechanisms that have any enforcement authority, whether within or outside of the UN system. Due to the sensitivities associated with proprietary technology, defining a governance mechanism that actors would trust becomes challenging.

[1] The current technological conversation largely revolves around generative AI technologies and systems. Generative AI may prove to be the most transformative – and potentially the most harmful – form of AI to date, if left without clear safeguards. However, we believe that soft law functions must encompass AI systems more broadly, also including more traditional, non-generative AI systems.

Short of hard law and enforcement, then, what can be done? An array of soft law functions can and should be fulfilled. While some of these functions are already being addressed through principles, standards, and policy recommendations endorsed by international bodies such as UNESCO,[2] OECD,[3] IEEE,[4] ISO,[5] and the EU[6], the landscape of international AI governance is fraught with inconsistencies, fragmentation, and an absence of critical functions. For example, there is no comprehensive international incident report or registry to track AI-related incidents, leaving states and organizations to operate in an informational vacuum. These omissions not only inhibit the proactive management of AI's global implications but also complicate the eventual transition to a hard law framework.

A serious need exists for international mechanisms to facilitate dialogue, cooperation, oversight, the development of effective confidence building measures, as well as verification and certification practices that can ensure safety, distribute benefits widely, and mitigate the risks and undesired societal consequences of AI. For example, not all states are able to undertake extensive review of AI applications they may consider implementing. An international body could help those states clarify which tools have undergone testing, compliance review, and certification, and understand any challenges and tradeoffs encountered by other countries that have deployed them.

Proposal

Most functions that can be specified as candidates for soft law are quite broad. They need to be analyzed and reduced to specific tasks that international governance can fulfill. In June and July 2023, The Carnegie Council for Ethics in International Affairs (CCEIA) and the Institute of Electrics and Electronics Engineers Standards Association (IEEE SA) hosted three expert workshops to take the first steps in this process. Our reports draw on ideas and insights from these workshops, and integrate them into a broader overview of soft law functions in the context of AI.

---

[2] Ramos, G. (2022). *Ethics of artificial intelligence*. UNESCO. Available at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

[3] OECD. (2019). *The OECD Artificial Intelligence (AI) Principles*. OECD. Available at: https://oecd.ai/en/ai-principles.

[4] IEEE Standards Association. *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.* (n.d.). IEEE Standards Association. Available at: https://standards.ieee.org/industry-connections/ec/autonomous-systems/.

[5] ISO. (2022). *ISO/IEC TR 24368:2022.* ISO. Available at: https://www.iso.org/standard/78507.html

[6] European Commission. (2021). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future.* European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

This paper is the second output from those workshops. The first was the *Framework for the International Governance of AI,*[7] presented in July to the UN AI Interagency Leadership Council and the ITU's AI for Good Summit in Geneva, and since distributed and discussed widely among policy makers. Both are part of a larger project to develop potential governmental structures for the international governance of AI.

The *Framework for the International Governance of AI* proposed five symbiotic components:
(1.) A neutral technical organization charged with continuously assessing which legal frameworks, best practices, and standards are achieving the highest levels of acceptance globally.

(2.) A normative governance capability with limited enforcement powers to promote compliance with global standards for the ethical and responsible use of AI and related technologies.

(3.) A toolbox for organizations to assess and certify conformity with standards.

(4.) The ongoing development of AI-governance-supporting technological tools, that can assist with data relevant for decision making, validating and auditing existing systems, and mitigating risks where necessary.

(5.) Creation of a Global AI Observatory (GAIO)[8], bridging the gap in understanding between scientists and policymakers and fulfilling the functions defined below that are not already being fulfilled by other institutions.

We propose that only component (2) requires an international treaty. All of the other components can be assembled to fulfill various soft law functions. In other words, this paper elucidates soft law functions in the context of international AI governance that do not require hard laws and regulations, along with the corresponding mechanisms necessary for their realization. The most important of these components is a GAIO, which, although it would benefit from being based on a treaty or, at a bare minimum, established through a resolution, could also be initiated through soft law measures.

We view these soft law functions as falling into three baskets: research directed at AI safety and scientific integrity; governance; and inclusivity (context and age appropriate) in design and decision making. Cutting across and uniting all three areas are communication and oversight. We expect that many of these tasks will be distributed across various existing institutions, from

---

[7] _____ (2023). *A Framework for the International Governance of AI.* Carnegie Council for Ethics in International Affairs. Available at: https://www.carnegiecouncil.org/media/article/a-framework-for-the-international-governance-of-ai.

[8] Mulgan, G., Malone, T., Siddharth, D., Huang, S., Tan, J., & Hammond, L. (2023). *The Case for a Global AI Observatory (GAIO), 2023.* Carnegie Council for Ethics in International Affairs. Available at: https://www.carnegiecouncil.org/media/article/the-case-for-a-global-ai-observatory-gaio-2023.

standard-setting bodies such as the IEEE and ISO to provisions of the EU's AI Act[9] that will be treated as *de facto* standards. For distributed governance to work effectively, however, a mechanism will also be needed that can drive cooperation and communication among stakeholders worldwide and implement the many disparate additional tasks that are required for effective oversight and not fulfilled by other institutions. Hence, this paper proposes the establishment of a GAIO that embodies this mechanism.

The below overview of soft law functions is not meant to be exhaustive. Needs may change as circumstances progress. The development of a new international body may be within the UN system, independent, or quasi-independent under the framework of the UN or another international institution. In this paper, we are agnostic about which approach will be taken – except to stress that whatever institution is charged with fulfilling these functions should represent the world as a whole, as AI is a scientific method and technology with global implications.

Definition of Soft Law

"Soft law" means non-binding norms, codes of conducts, principles, standards, or guidelines that lack the enforceable character of hard law. This paper uses the term broadly to encompass any mechanism that helps direct the deployment of AI systems toward ethically desirable goals. Requirements for insurance coverage, laboratory practices and procedures, and adherence to sound ethical principles as a prerequisite for the publication of research are all examples of soft law.

Even without the legally binding obligations of treaty frameworks, soft law can entail non-legal enforcement and often plays a crucial role in shaping state behavior and setting normative expectations.[10] It significantly contributes to the evolution of international norms, customary practices and global governance. Instruments of soft law can be alternatives to treaties, or can serve to complement, clarify, or amplify a treaty. They can be precursors of treaties, allowing states to test out commitments before formalizing them: the Universal Declaration of Human Rights,[11] for example, was initially adopted without legally binding force.

---

[9] European Parliament. (2023). *EU AI Act: First Regulation on Artificial Intelligence | News | European Parliament*. European Parliament. Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[10] Gutierrez, C. I., Marchant, G., & Kaspersen, A. (2021). *Soft Law Approaches to AI Governance.* Carnegie Council for Ethics in International Affairs. Available at: https://www.carnegiecouncil.org/media/series/aiei/20210707-soft-law-artificial-intelligence-governance

[11] United Nations. (1948). *Universal Declaration of Human Rights.* United Nations. Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

For the purpose of this paper, *soft law functions* describe high-level objectives to support the responsible development and deployment of AI models. They can be fulfilled by mechanisms put in place for the oversight and governance of AI without a multilateral treaty. These functions and mechanisms can be broad in scope and may or may not culminate in explicit standards or guidelines. A specific soft law can refer to a clear task or requirement, while a *soft law function* may encompass a wide array of standards and practices. Whether or not these mechanisms lead to a treaty is less important than their role in facilitating stakeholder interactions. To be sure, a treaty would be preferable – but we propose that any meaningful discussion about AI international governance should consider that soft law mechanisms are put in place in advance of a treaty, as they can be built on if and when a treaty is agreed. We recognize that some states are dismissive of soft law considering it to be insufficient or weak. Given the serious societal and ethical challenges AI poses for individual states and for worldwide security, stability, and equity, we believe the establishment of international mechanisms for AI governance can not wait for a treaty.

Soft Law Functions

Both hard and soft laws have three components: a function, mechanisms, and a body that executes the mechanisms to fulfill the function. The following sections present an overview of soft law functions and mechanisms for AI governance. The functions we outline may not all lead directly to soft law provisions, but they are each essential for formulating the areas in which soft law, and eventually hard law, will be necessary. These functions are not mutually exclusive. International dialogue can, for example, also benefit expectation setting; however, we decided to include them as separate functions since they serve distinct purposes in the context of international AI governance.

International Cooperation and Dialogue

International challenges require collective solutions. Fostering dialogue among nations and stakeholders can pave the way for more substantial, sometimes even legally binding, commitments and confidence building measures.

The core mechanism of this function is driving communication, cooperation, collaboration – and where possible, coordination. Evaluating existing soft law and considering additional needs can be a catalyst for countries and international bodies to communicate. For instance, international conferences or forums where states discuss common challenges often rest on soft law principles. They can be a venue to air differences, build trust and confidence, encourage transparency, and invite a diversity of views and experiences. By building shared understanding, they can set the

stage for multilateral initiatives or collaborative approaches to global issues, such as mitigating harmful applications of AI systems and distributing their benefits.

Non-binding resolutions and declarations, such as those from the UN General Assembly,[12] can also lay the groundwork for enhanced cooperation by fostering consensus and setting aspirational goals. By signaling the global community's commitment to addressing collective challenges, these declarations carry significant normative weight.
Central to this ambition is furthering a multidisciplinary, multistakeholder, transnational dialogue and scientific collaboration on the different impacts of AI across borders, cultures and communities. Consultations framed around civic engagement can foster dialogue among scientists, engineers, developers, policymakers, and social scientists on how to ensure that AI developments are not only technically robust but also context sensitive, age appropriate, ethically aligned and socially beneficial. While there are some initiatives that have attempted to promote international cooperation and dialogue, such as the Partnership on AI[13] and the Global Partnerships on AI[14], these initiatives are primarily controlled by leading corporations producing AI applications and the interests of the G7 respectively. They have not progressed significantly beyond expressing verbal support for the necessity of international AI governance.

Understanding Opportunities and Risks of AI Systems

Building a shared understanding among stakeholders on the opportunities and the potential harmful uses of AI, is a step towards creating a unified approach to harnessing the benefits while mitigating the risks of the technology; both aspects are significantly understudied and policymakers are acting in an information vacuum. Either that, or they are largely dependent upon work performed by corporations leading the deployment of AI, and whose actions are primarily directed by their fiduciary responsibility to shareholders. The current AI landscape lacks a centralized institution that can offer insights into opportunities and risks, with much of the work being kept confidential. The proposed GAIO could bridge this knowledge gap, mirroring the role of other intergovernmental mechanisms that provide governments with timely and independent scientific research and expertise.

Hence, research into potential harms and benefits is the primary mechanism to fulfill this function, allowing for forward-looking, proactive policy development. Beyond safety and ethical principles, the Sustainable Development Goals[15] (SDGs) provide one possible framework for

[12] See, for example, resolutions 72/242 and 73/17 in which the General Assembly acknowledges that swift and extensive technological advancements can greatly influence sustainable development in both beneficial and adverse ways. To capitalize on the benefits and tackle the challenges, such international collaboration involving multiple stakeholders is essential.
[13] For more information on the Partnership on AI, see: https://partnershiponai.org/
[14] For more information on the Global Partnership on AI, see: https://gpai.ai/
[15] United Nations. (2015). *The 17 sustainable development goals.* United Nations. https://sdgs.un.org/goals

research to assess the societal impact of AI systems across different dimensions. For example, AI could help to advance the SDGs on healthcare and education while setting back the SDG on reducing inequalities.[16]

Such research must consider the characteristics of AI systems that might induce harmful impact. Pinpointing specific features or behaviors, as well as identifying use cases, that lead to undesirable outcomes makes it possible to design policy instruments that make the development of AI systems inherently safer and more reliable. For example, inherent biases in databases is one known characteristic that leads to gender and racial biases in output. Unfortunately, ameliorating even known characteristics that can be harmful such biases may be quite difficult.

In the AI risk research community, a divide has emerged between what are often referred to as short-term and long-term risks[17]. This binary perspective, which applies to both comprehending the limitations of AI systems and the constraints of risk-based approaches, has hindered informed public discourse on these matters. Consequently, there is an urgent need for an organization that can comprehensively and impartially assess all risks, with a particular emphasis on the safety and security of critical systems embedded into public infrastructure. This entity will be crucial in providing independent analysis and support for informing political deliberations regarding the governance and necessary oversight of AI technology. It would enable policymakers to grasp the relevance and immediacy and respond in a timely fashion to potential technological threats.

Expectation Setting and Harmonization

Establishing clear expectations can foster a shared understanding among developers, users, and regulators on aligning AI systems with defined safeguards, societal values and ethical standards. This approach promotes proactive responsibility, allowing the AI community to anticipate challenges and address them collaboratively.

Expectation setting will, on the one hand, depend on overcoming the current fragmentation in policy approaches towards AI governance, with numerous sets of principles wrapped in national agendas – such as Canada's Guiding Principles for AI,[18] Australia's AI Ethics Principles,[19] and

[16] The International Research Centre on AI and UNESCO have released a report listing the 100 best uses of AI to solve specific problems covered by the SDGs. Available at: https://ircai.org/global-top-100-outstanding-projects/results/

[17] Short-term risks refer to bias, discrimination, and misinformation, for example, while long-term risks mean, for example, potential threats leading to human extinction by advanced general intelligence.

[18] Government of Canada. (2018). *Responsible use of artificial intelligence (AI).* Government of Canada. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#.

[19] Department of Industry, Science and Resources. (2022). *Australia's AI Ethics Principles.* Government of Australia. Available at: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles.

the Beijing Artificial Intelligence Principles.[20] For example, while the Beijing Artificial Intelligence Principles share similarities with others, they replace the term "human rights" with “harmony", reflecting a Chinese cultural perspective on societal interactions. The operationalization of such differing sets of principles may be unproblematic within states, but can give rise to confusion, disagreements and diverging expectations once applications are deployed outside of national borders.

Some states have started to draft codes of conduct specifically directed at the development and deployment of AI systems across international borders to set shared expectations. For example, the EU-US Trade Technology Council has been working on a voluntary, bilateral code of conduct[21] that emphasizes transatlantic cooperation on transparency, risk audits, and other technical details.

On the other hand, professional technical organizations, and NGOs are working to translate principles and recommendations into actionable standards. The AI standards developed by ISO , for example, can serve as valuable reference points, e.g., on machine learning[22] and trustworthiness.[23] The IEEE has also formulated standards to address the ethical implications of AI in areas such as generative pretrained models,[24] emulated empathy,[25] and age appropriateness.[26]

While the implementation of standards through certification, conformity assessments, research collaboration, and scientific publications can contribute to setting expectations, a general concern with standards development is the limited involvement of civil society, and dominance of

---

[20] International Research Center for AI Ethics and Governance. (2022). *Beijing Artificial Intelligence Principles.* International Research Center for AI Ethics and Governance. Available at: https://ai-ethics-and-governance.institute/beijing-artificial-intelligence-principles/#:~:text=Human%20privacy%2C%20dignity%2C%20freedom%2C,utilize%20or%20harm%20human%20beings.

[21] The White House. (2023). *U.S.-EU Joint Statement of the Trade and Technology Council.* The White House. Available at: https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/31/u-s-eu-joint-statement-of-the-trade-and-technology-council-2/

[22] ISO/IEC JTC 1/SC 42 Artificial intelligence Committee. (2022). *ISO/IEC 23053:2022.* ISO. Available at: https://www.iso.org/standard/74438.html.

[23] ISO/IEC JTC 1/SC 42 Artificial intelligence Committee. (2020). ISO/IEC TR 24028:2020. ISO. Available at: https://www.iso.org/standard/77608.html.

[24] Generative Pretrained AI Models Working Group. (2023). *P7018 - Standard for Security and Trustworthiness Requirements in Generative Pretrained Artificial Intelligence (AI) Models.* IEEE Standards Association. Available at: https://standards.ieee.org/ieee/7018/11306/

[25] Empathic Technology Working Group. (2019). *P7014 - Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems.* IEEE Standards Association. Available at: https://standards.ieee.org/ieee/7014/7648/.

[26] Age Appropriate Digital Services Framework Working Group. (2021). *IEEE 2089-2021 - IEEE Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children.* IEEE Standards Association. Available at: https://standards.ieee.org/ieee/2089/7633/.

industry players and government agendas. Some organizations such as IEEE are addressing this issue through governance structures; for example, their standards formulating committees do not include state parties and consist solely of experts; an intricate balloting system inhibits agenda capture by vested interests. This can encourage academics, policy planners, and other stakeholders to join committees formulating standards to ensure diversity of views to inform standards and subsequently shape expectations.

Finally, supporting the harmonization of standards, both from private and public entities, is a mechanism that is necessary for this soft law function. Promoting the adoption of shared standards and best practices can create a more unified global approach to AI governance[27]. There is significant role to be played by a new UN-supported mechanism in guiding this process, given the need for global cooperation in legitimizing shared standards and working through differences.

Provision of Technical Tools Supporting AI Governance

Due to their complexity and potential impact, AI systems are at risk of being misused and potential use-cases at risk of being ill suited, whether intentionally or accidentally. For instance, generative AI-based synthetic disinformation has negatively impacted trust in the internet and society, posing threats to global stability.[28] Tools to detect, track and trace back such disinformation would be useful to support governance initiatives tackling the issue.

The role of developing technological tools in supporting AI governance is often relegated to industry actors, a situation that leaves gaps in addressing public interest concerns that are not immediately profitable or attractive for corporate initiatives. While industry-driven technological solutions are undoubtedly valuable, they are not comprehensive in their reach, often failing to address issues that do not align with market incentives. For example, the development of tools to detect and mitigate AI-generated disinformation may not be commercially appealing, but are essential for effective governance interventions.

This underlines the critical need for an international body that takes up the slack by promoting the development of such tools. In instances where the industry is reluctant to develop necessary tools, the governance body could stimulate their development through, for example, financial incentives or direct governmental involvement in their creation.

---

[27] Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., Ó hÉigeartaigh, S., Staffell, S., Villalobos, J. (2023). *International Governance of Civilian AI: A Jurisdictional Certification Approach*. Available at: https://www.oxfordmartin.ox.ac.uk/publications/international-governance-of-civilian-ai-a-jurisdictional-certification-approach/.

[28] Banias, M. J. (2023). *Inside CounterCloud: A Fully Autonomous AI Disinformation System.* The Debrief. Available at: https://thedebrief.org/countercloud-ai-disinformation/

Informing Diplomats, Legislators, and the General Public

Neutral, evidence-based information on the state of AI is vital to inform decision-making and foster transparency, accountability, and trust among stakeholders, from policymakers to the general public.[29]

Four registries have been proposed in this context[30] to function as neutral information repositories with the aim of documenting a wide range of AI-related information and synthesize evidence to support diverse governance responses: First, a registry of adverse incidents would provide insights into challenges and risks of AI technologies[31]. Second, a registry of emerging and anticipated AI applications would enable stakeholders to prepare for future developments. Third, a registry chronicling the history of AI systems – detailing testing, verification, updates, and experiences of states that have deployed them – would aid countries that lack the resources to evaluate systems, and ensure that lessons from past deployments inform future actions. The final registry would maintain a global repository for data, code, and model provenance. These repositories would further support identifying and documenting best practices on a technological level. By giving stakeholders access to the latest and most effective strategies in the development and deployment of AI systems, common pitfalls and mistakes can be avoided.

Besides these registries, standardized reporting is another essential mechanism. Reporting AI-related incidents, developments, and insights in a consistent manner facilitates comparability and analysis by ensuring that stakeholders can access, understand, and act on the information provided. Reports should further be as close to real-time as possible to facilitate early responses that can prevent or minimize harm.

The OECD has created an AI Policy Observatory, which in its current form has different goals and fulfills different functions than most of those described in this article for a Global AI Observatory (GAIO). Furthermore, the OECD does not represent all countries and stakeholders. However, the existence of the OECD AI Policy Observatory has created some confusion with the proposal for a new GAIO. Rather than a GAIO this collection of tasks could be labeled with a different name. Previous proposals for something similar have been called a Governance

[29] While difficult to build trust for stakeholders to share these information, setting up a trustworthy apparatus is necessary. We know this aspect is central and it's part of our broader project but not germane to this paper.
[30] Mulgan, G., Malone, T., Siddharth, D., Huang, S., Tan, J., & Hammond, L. (2023). *The Case for a Global AI Observatory (GAIO), 2023.* Carnegie Council for Ethics in International Affairs. Available at: https://www.carnegiecouncil.org/media/article/the-case-for-a-global-ai-observatory-gaio-2023.
[31] This would be similar to the AI Incident Database (https://incidentdatabase.ai/) but more comprehensive, including incidents that have not been publically reported on and those that occurred outside of industry.

Coordination Committee[32], Global Governance Coordinating Committee[33], AI Global Governance Network[34], and a Network of Networks[35].

Finally, in addition to voluntary confidence-building measures such as self-reporting, it may be prudent to consider implementing whistleblowing channels to assist with the neutral provision of information. These channels would offer a safe and confidential way for individuals to report unethical, illegal, or harmful activities related to AI. By empowering individuals to come forward with information, we increase the likelihood of promptly identifying and addressing potential risks or malpractices. Such channels are particularly crucial in fields like AI, where the rapid pace of advances and the proprietary nature of technologies can sometimes obscure what is at stake or enable unethical practices.

Policy Design and Implementation Assistance

This function bridges the technical intricacies of AI and the national legal and regulatory frameworks that seek to guide its application. The objective is to ensure that laws and regulations not only respond to the current state of AI technology but also look forward, anticipating future developments and challenges. UNESCO, for example, is working to help individual states to implement its AI principles and policy proposals.

The first mechanism to fulfill this function is the evaluation of legislative and regulatory approaches. Assessing the pros and cons of each option enables policymakers to make more informed decisions that strike the right balance between fostering innovation and ensuring public safety and ethical considerations. This process helps to identify gaps in existing regulations, potential overlaps, and areas where new legislative interventions might be necessary.

The second mechanism is *model governance*, offering a blueprint for AI governance at the state or regional level. States, of course, differ in their needs, desires, and capacity to build national governance infrastructure for AI. Nevertheless, models could serve as reference points that allow individual states to tailor their regulations to their particular customs and needs, while helping to maintain a core set of governance principles and standards. Providing a standardized framework can also help to achieve a more consistent approach to AI regulation across jurisdictions. This

---

[32] Marchant, G. E. & Wallach, W. (2015). *Coordinating Technology Governance*. Issues in Science & Technology, 31[4] pp. 43–50, Summer 2015.

[33] W. Wallach and G.E. Marchant (2017). *An Agile Ethical/Legal Model for the International and National Governance of AI and Robotics*. The Hastings Center.

[34] Proposal for an *AI Global Governance Network* prepared for the International Congress for the Governance of AI, which was to have met in Prague, Czech Republic in May 2020, but was canceled due to the covid epidemic.

[35] Slaughter A. & Chehadi, F. (2023). *AI's Pugwash Moment*. Project Syndicate. Available at: https://www.project-syndicate.org/commentary/institutions-to-govern-artificial-intelligence-new-pugwash-movement-by-anne-marie-slaughter-and-fadi-chehade-2023-07?barrier=accesspaylog

facilitates smooth inter-state collaborations and transactions, and provides businesses and developers with a clear set of guidelines to adhere to, wherever they operate. Compared to regular standard setting, which focuses on disjoint technical guidance, model governance would be more comprehensive in nature, ensuring that all governance components are effectively tied together. Model governance could be complemented by the provision of a *governance toolkit* – a practical guide for stakeholders involved in AI development and deployment. This toolkit could provide resources, technical assistance, best practices, and guidelines across various sectors, approaches, and applications. From data privacy to ethical considerations or technical standards, it could offer actionable insights that can be adapted to specific contexts.

Evaluation and Monitoring

Monitoring and evaluation are not merely about oversight; they are a proactive approach to ensure that AI systems operate within defined ethical, legal, and technical boundaries. This function aims to track the development, deployment, and impact of AI, and its alignment with societal values and standards.[36]

The first mechanism in this context are conformity assessments and certifications of AI systems[37]. This process involves evaluating systems against established benchmarks[38] or standards to ensure their reliability, safety, and ethical soundness. It encompasses translating identified best practices into specific organization- and system-level requirements. Certifying AI systems that meet defined criteria signals to users, developers, and regulators that the system has been rigorously tested and deemed fit for its intended application. This not only builds trust in AI applications but also provides a clear framework for developers to design their systems.

The second mechanism is often referred to as compute monitoring[39] – tracking the tangible and intangible resources that AI systems utilize. As models become more complex, they demand more computational power, data, and other resources. Monitoring these requirements helps to understand the environmental impact of AI, given the energy-intensive nature of some computations; identify if resources are being monopolized; and provide insights into the scalability and sustainability of AI applications. All this helps support targeted AI governance initiatives on a national and international level.

---

[36] Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., & Snidal, D. (2023). *International Institutions for Advanced AI.* Available at: https://arxiv.org/pdf/2307.04699.pdf.

[37] One example is the certification program by The Responsible AI Institute. However, most of the programs currently available aren't providing their certification guidelines publicly, which prevents public scrutiny of and contribution to the underlying assessment methodology.

[38] While there exist some benchmarks for AI systems, they aren't fully established yet nor are they reliably and comprehensively assessing key responsibility characteristics of AI systems.

[39] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning.* Available at: https://arxiv.org/pdf/2202.05924.pdf?trk=public_post_comment-text.

Encouraging Development of and Accessibility to Beneficial Technology

AI can and should be a tool for societal advancement. Its benefits should be widespread, not confined to a privileged few. Compared to the provision of technical tools to support AI governance processes and initiatives, this function ensures that the technology can be accessed by the public and is used to solve pressing societal challenges.

One mechanism to achieve this is developing cutting-edge AI models for the public good[40] in areas such as healthcare, environmental conservation, or education where resources would otherwise be insufficient to build these models in contexts where commercial interest is absent. Such models, when made open-source or easily accessible, can act as a foundation upon which various stakeholders, including researchers, NGOs, and governments, can build solutions tailored to their specific contexts.

The second mechanism is building the infrastructure that makes the development and deployment of AI possible. Investing in hardware makes more computational power available to develop and run sophisticated AI models, potentially lowering barriers to entry and fostering innovation and inclusivity. This includes digital infrastructure, such as the internet, and physical infrastructure, such as hardware. Widespread internet access democratizes the availability of AI-powered solutions, allowing even remote and underserved communities to benefit from technological advances. According to a 2021 report from the ITU, 2.9 billion people worldwide are without internet access, 96% of them in developing countries.[41]

Building Public Trust in Institutional Oversight of AI

As AI technologies become increasingly embedded in our daily lives, the public must have confidence in the ability of the institutions that oversee these technologies to ensure that benefits are realized and potential harms are mitigated and that companies are being held accountable. Shifting risk onto the users of AI applications, or those impacted by their premature deployment, is not responsible governance, good business practice, or conducive to establishing trust and confidence.

One mechanism to foster public trust and confidence is the publication of an annual report on the state of AI and AI governance[42]. This report would provide a comprehensive overview of AI

[40] Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., Snidal, D., & Deepmind, G. (2023). *International Institutions for Advanced AI.* Available at: https://arxiv.org/pdf/2307.04699.pdf.
[41] ITU. (2021). *Facts and Figures 2021: 2.9 billion people still offline.* ITU. Available at: https://www.itu.int/hub/2021/11/facts-and-figures-2021-2-9-billion-people-still-offline/
[42] As recommended by G. Mulgan et al. (2023) as a function to be fulfilled by the GIAO.

advancements, highlight challenges in design and development, and outline the steps taken by all actors and institutions to address these challenges. Offering a transparent account of the AI landscape and governance measures would reassure the public that oversight bodies are vigilant and adapting to the ever-evolving world of AI.

As part of any confidence-building regime, periodic interim reports (self-reporting) should be issued throughout the year. Given that a year can bring significant developments in the field of AI, these interim reports would address any noteworthy developments that arise between the annual reviews, such as breakthrough research, notable incidents, or changes in governance policies. These reports would keep the public informed in real-time, reinforcing the commitment of institutions to transparency and responsiveness.

Easily understandable information, free from hype and grounded in scientific integrity, is vital for establishing public trust. While detailed reports are necessary for a comprehensive understanding, they can be overwhelming for the average citizen. Thus, there is a need for simplified, clear, and accessible information that allows everyone to grasp the essence of policy decisions, understand the rationale behind them, and comprehend their implications.

Advocacy and Inclusivity

The governance structures overseeing AI must be representative of the diversity of humanity. Advocacy, in this context, is not just about promoting a particular viewpoint but ensuring inclusivity, equity, and justice in the development and deployment of AI. While this function could be subsumed under ‘International Collaboration and Dialogue’, we decided that given it’s neglectedness in the current ecosystem and its necessity for effective international AI governance efforts, it is justified to highlight it as a separate function.

The impact of AI technologies is felt across borders, cultures, and communities, but the discourse around AI governance has often been dominated by a select few, primarily from technologically advanced regions. Central to this advocacy function is the mechanism of actively including in international AI governance and decision-making processes groups that have often been marginalized in the global discourse, whether due to geography, socio-economic status, gender, ethnicity, or other factors.

It is not just ethically right but pragmatically essential to tap into a wealth of diverse perspectives, experiences, and insights. Solutions and policies born out of a diverse deliberative process are more likely to be holistic and robust, addressing potential blind spots that a more homogenous group might overlook. When underrepresented groups see themselves as active participants in governance, it also fosters a sense of ownership and trust in the AI systems that permeate their lives.

Conclusion

The speed at which AI systems and technologies are being developed and deployed – and the range of misuses, abuses, and undesirable societal consequences for which they can be adapted – cries out for an international hard law regime with enforcement authority. While we hope the international community will act soon, past experience suggests that the pathway to put in place effective oversight and enforcement will be slow and laborious.

In the meantime, we are forced to rely on soft law mechanisms to help shape the trajectory of AI development and encourage international cooperation in its normative and technical governance. This paper outlines the international soft law functions for AI and the mechanisms needed to fulfill them. We must begin implementing these mechanisms now, assuming that they are not being addressed by existing institutions.In particular, we call for the immediate creation of a GAIO, in line with the proposal by Mulgan et al. (2023), which should initially undertake six areas of activity to fulfill functions outlined above:

- Maintain a global database for standardized reporting of incidents with real-world consequences – for example, use of AI to create a dangerous pathogen. This would support cross-border coordination to mitigate emerging threats.
- Maintain a registry of AI systems with the largest social and economic impacts, and track those impacts. Some governments have started work on such systems at national level, but a global approach would be more effective.
- Assemble data and conduct analysis on facts related to AI, such as levels of investment, geography, uses, and applications. There are many sources for these data, but they are not brought together in an easily accessible form.
- Convene working groups to assess the positive and negative impacts of AI on areas such as labor markets, education, media, and healthcare. These groups would gather and interpret data and make forecasts on potential future effects.
- Develop models for regulations, laws, and policies, and offer national governments assistance in adapting those models to their particular contexts. This work would draw on lessons from Co-develop promoting DPI and IAEA.
- Publish an annual report that summarizes emerging patterns, outlines scenarios for the coming two to three years, and set out choices for governments and international organizations.[43]

To be sure, the structure of a GAIO and a roadmap for its implementation must still be developed. Furthermore, developing structures and institutions that can instill confidence , in a landscape dominated by diverging agendas and interests, will not be easy. Nevertheless, there is

[43] This could build on efforts such as Stanford's AI Index, see https://aiindex.stanford.edu/report/.

a growing chorus in favor of establishing an international body to coordinate the activities of institutions that fulfill some of the described functions and take on responsibilities currently unaddressed or addressed insufficiently by other institutions. This step is necessary to ensure the responsible global development and deployment of AI.

Should the demand arise, the GAIO and other governance mechanisms developed can be built upon as foundations to speed up the implementation of treaties adopted and the enforcement of international hard law obligations. As international governance moves toward establishing effective international oversight mechanisms and frameworks, the models in place will likely serve broader governance needs – including, in the future, for other emerging technologies not yet realized.