Soft Law Functions in the International Governance of Al

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 and equity – as well as political, social, 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, i.e., objectives in the context of international AI governance, and mechanisms to fulfill them. We further propose the establishment of a global AI governance mechanism to fulfill AI soft law functions that have not been (sufficiently) picked up by existing institutions.

Introduction

Recent calls for the international governance of AI¹ seldom include any specifics. A few governance initiatives are under development that have international ramifications, but they are driven by small groups of states or are largely conceptual². 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. Given the sensitivities related to proprietary technology, it is also difficult to define any governance mechanism that states would trust.

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 bodies such as UNESCO,³

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 target AI systems more broadly, also including more traditional, non-generative AI systems.

² Trager, R., Harack, B., Reuel, A., Carnegie, A., Heim, L., Ho, L., Kreps, S., Lall, R., Larter, O., hÉigeartaigh, S., Staffell, S. (2023). *International Governance of Civilian AI: A Jurisdictional Certification Approach*. Available at: https://arxiv.org/pdf/2308.15514.pdf

³ Ramos, G. (2022). *Ethics of artificial intelligence*. UNESCO. Available at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

OECD,⁴ IEEE,⁵ ISO,⁶ and the EU⁷, 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 communication, cooperation, oversight, and the development of effective strategies 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 reviews 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.

Many activities that can be specified as candidates for international soft law are quite broad in scope. They need to be analyzed and reduced to specific goals 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 article draws on ideas and insights from these workshops, and integrates them into a broader overview of goals or soft law functions in the context of AI.

This paper is the second output from those workshops. The first was the *Framework for the International Governance of AI*, ⁸ 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 makers⁹. It and accompanying documents are part of a larger initiative to develop potential governmental structures for the international governance of AI.

⁴ OECD. (2019). *The OECD Artificial Intelligence (AI) Principles*. OECD. Available at: https://oecd.ai/en/ai-principles.

⁵ 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/. ⁶ ISO. (2022). *ISO/IEC TR 24368:2022*. ISO. Available at: https://www.iso.org/standard/78507.html

⁷ 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.

⁸ _____ (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.

⁹ The framework was also updated to include possible modalities and «middleware» and submitted to the UN Secretary Generals call for papers on International AI Governance. Available at: https://www.carnegiecouncil.org/media/article/envisioning-modalities-ai-governance-tech-envoy

The Distinction between Soft Law & Soft Law Functions

"Soft law" refers to 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 desirable goals. Requirements for insurance coverage, responsible scaling practices ¹⁰ and safety 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. ¹¹ It can significantly contribute to the evolution of international norms, customary practices and global governance. Instruments of soft law can, in some instances, be normative 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, ¹² for example, was initially adopted without legally binding force.

For the purpose of this paper, *soft law functions* describe high-level objectives or goals to support the responsible development and deployment of AI models. Given the breadth of specific soft law activities that AI systems require or might generate, each of these functions will encompass a broad range of soft laws including, for example, standards and best practices.

International soft law functions can be fulfilled by mechanisms put in place for the oversight and governance of AI without a multilateral treaty. A specific soft law mechanism can refer to a clear task or requirement, while a soft law function is a goal that may be achieved by a wide array of soft law mechanisms.

The potential of these mechanisms to lead to a treaty should be considered separately from their essential role in facilitating stakeholder interactions. It is clear that a treaty would be preferable; however, it is vital to recognize that meaningful discussions about AI international governance must prioritize the establishment of soft law mechanisms beforehand. These mechanisms can serve as a foundational framework that can be built upon and enhanced with clear enforcement capacities if and when a treaty is agreed upon.

¹⁰ Anthropic. (2023). *Anthropic's Responsible Scaling Policy*. Available at: https://www.anthropic.com/index/anthropics-responsible-scaling-policy

¹¹ 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

¹² United Nations. (1948). *Universal Declaration of Human Rights*. United Nations. Available at: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

We acknowledge that some states are skeptical of soft law, viewing it as inadequate, weak, or merely establishing a normative framework without necessitating national concessions. However, considering the profound societal, ethical, and security challenges AI poses both to individual states and global stability and equity, the urgency of establishing international mechanisms for AI governance cannot be overstated. These mechanisms are crucial and should not be delayed in anticipation of a treaty. This overview of soft law functions is intended to be illustrative rather than exhaustive, acknowledging that requirements may evolve as circumstances change. The creation of a new international body to carry out these functions could be within the UN system, operate independently, or function as a quasi-independent entity under the aegis of the UN or another international institution. While we remain open-minded in this paper about the specific approach to be adopted, it is imperative to emphasize that any institution tasked with these responsibilities must be representative of the global community, especially considering AI's extensive and tangible impact on areas such as global security, economic development, and international political, social, and human rights.

Overview of Soft Law Functions

Both hard and soft laws entail three key elements: their function, the mechanisms employed, and the executing body responsible for these mechanisms to achieve the intended function. This paper offers an overview of the soft law functions and corresponding mechanisms pertinent to AI governance utilizing soft law. These functions are interrelated, yet we treat them as distinct functions due to their unique roles in the realm of international AI governance. All of the components are essential for formulating the areas in which soft law, and in most cases eventually hard law, will be necessary.

Understanding Opportunities and Risks of Al Systems

Building a shared understanding among stakeholders about the opportunities and risks of AI is crucial for developing a cohesive strategy to maximize its benefits and minimize its risks. Alas, at present, the information landscape regarding AI is often fragmented or skewed, aligning more with narrow industry or national interests than with a comprehensive understanding. This leads to a situation where decision-makers navigate in a realm shaped by information realities that primarily benefit these interests. As a result, they frequently depend on research either conducted by or financed by the very corporations that dominate AI design, development, and deployment.

These corporations' actions and decisions are primarily influenced by their fiduciary responsibilities to shareholders and their drive for market dominance in a data-driven economy. The current state of AI lacks a centralized authority that can offer unbiased insights into the technology's potential opportunities and risks. The paucity of transparent and impartial research renders the true nature and impact of AI systems as somewhat clandestine. There's a pressing

need for a body similar to other intergovernmental mechanisms, providing governments with timely, independent scientific research and analysis.

Hence, research into both current and potential future risks and benefits of AI serves as the primary mechanism to fulfill this function. Such research enables effective, forward-looking, and proactive policy development. Beyond safety and ethical principles, the Sustainable Development Goals (SDGs)¹³ offer a comprehensive framework for assessing the societal impacts of AI systems across various dimensions. For instance, AI has the potential to significantly advance SDGs related to healthcare, well-being, economic growth and education, but it could also inadvertently hinder the availability of jobs and exacerbate inequalities, as reflected in the SDGs on decent work and reducing disparities. Furthermore, the use of AI in promoting peace, justice and strong institutions is another crucial area, underscoring the need for an informed and balanced approach in AI deployment to ensure it aligns with and supports broader humanitarian and peacekeeping goals.

This research needs to thoroughly examine the characteristics of AI systems that could potentially cause harmful impact¹⁴. By identifying specific features or behaviors that result in undesirable outcomes, especially those with transnational implications, it becomes feasible to craft policy measures and modalities that enhance the safety and reliability of AI systems. For instance, inherent biases in databases are a well documented issue in these systems, leading to biased and discriminatory output. Unfortunately, ameliorating such harmful characteristics can be challenging.¹⁵

In the AI risks research community, a notable divide exists between what are termed as short-and long-term risks ¹⁶. Regardless of one's stance on these issues, there is currently a lack of an entity that assesses all risks in a comprehensive and impartial manner. This absence allows for the emergence of curated narratives that potentially exaggerate or understate the significance of certain risks, often aligning with the interests of those who stand to benefit most from emphasizing specific dangers over others. What is urgently required is an independent body that can provide policymakers with unbiased guidance on the relevance and urgency of the potential dangers associated with AI.

¹³ Available at: https://sdgs.un.org/goals

¹⁴ Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N. and Ho, L. (2023). *Model Evaluation for Extreme Risks*. Available at: https://arxiv.org/pdf/2305.15324.pdf

¹⁵ Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). *A survey on bias and fairness in machine learning*. ACM computing surveys (CSUR) 54, no. 6 (2021): 1-35.

¹⁶ Short-term risks refer to bias, discrimination, and misinformation, for example, while long-term risks mean, for example, potential threats to human extinction by advanced artificial general intelligence.

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 guardrails, transparent values and carefully thought through 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,¹⁷ Australia's AI Ethics Principles,¹⁸ and the Beijing Artificial Intelligence Principles.¹⁹ 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 conduct²⁰ that emphasizes transatlantic cooperation on transparency, risk audits, and other technical details.

On the other hand, professional technical organizations and NGOs are actively translating principles and recommendations into actionable standards. For instance, the AI standards developed by ISO can serve as valuable reference points, particularly in areas like machine

¹⁷ 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#.

¹⁸ 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.

¹⁹ 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-

 $principles/\#:\sim: text=Human\%20 privacy\%2C\%20 dignity\%2C\%20 freedom\%2C, utilize\%20 or\%20 harm\%20 human\%20 beings.$

²⁰ 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/

learning²¹ and trustworthiness²². Additionally, the IEEE SA has developed certification programs and formulated standards to address ethical concerns during the design of AI systems. These standards are crucial for grappling with the ethical implications of deploying AI in various fields, including generative pre-trained models²³, emulated empathy²⁴, and ensuring careful calibration for age appropriateness²⁵.

While the implementation of standards through certification, conformity assessments, research collaboration, and scientific publications can contribute to setting expectations, a general concern with most standards development processes is the dominance of industry players and government agendas, and in the case of nationalized standards bodies, the limited involvement of civil society. Some organizations such as the IEEE SA 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 governance²⁶. This requires overcoming competition among standard setting bodies, which earn revenue when their standards are adopted. There is a potential role for the UN or another international institution in this process, given the need for global cooperation in legitimizing shared standards and working through differences.

-

²¹ ISO/IEC JTC 1/SC 42 Artificial intelligence Committee. (2022). *ISO/IEC 23053:2022*. ISO. Available at: https://www.iso.org/standard/74438.html.

²² 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.

²³ 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/

²⁴ 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/.

²⁵ 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/.

²⁶ 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/.

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.

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 Assembly²⁷, are crucial for laying the foundation for improved cooperation. They do this by fostering consensus and setting aspirational goals. These documents often carry significant moral weight by signaling a commitment from the global community to tackle collective challenges. This is especially apparent in the realm of international arms control.

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 been trying to foster this international cooperation and dialogue, such as the Partnership on AI²⁸ and the Global Partnerships on AI²⁹, these initiatives are dominated by leading corporations producing AI applications and by members of the G7 respectively, rather than being inclusive to the world community, and have not moved significantly beyond the stage of verbalizing support for the need of international AI governance.

²⁷ 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.

²⁸ For more information on the Partnership on AI, see: https://partnershiponai.org/

²⁹ For more information on the Global Partnership on AI, see: https://gpai.ai/

Provision of Technical Tools Supporting Al Governance

Due to their complexity and potential impact, AI systems are at risk of being misused, 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.³⁰ Tools to detect, track and trace back such disinformation are needed 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 such as protecting consumers..

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.

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.³¹

Mulgan et al. (2023) proposed four registries in this context³² 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 technologies³³. Second, a registry of emerging and anticipated AI applications would enable stakeholders to prepare for future developments.

³⁰ Banias, M. J. (2023). *Inside CounterCloud: A Fully Autonomous AI Disinformation System*. The Debrief. Available at: https://thedebrief.org/countercloud-ai-disinformation/

³¹ 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.

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.

³³ 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.

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 as well as the first soft law function we mentioned, understanding opportunities and risks of AI. 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 for this soft law function. 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.

Finally, whistleblowing channels assist with neutral information provision by providing 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, they make it more likely that potential risks or malpractices will be promptly identified and addressed. Such channels are especially vital in fields such as AI, where the speed of advances and the proprietary nature of technologies can sometimes obscure challenges or unethical practices.

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 compliance of AI systems with best practices and safety standards and to evaluate the capabilities of individual systems.³⁴

The first mechanism in this context are conformity assessments and certifications of AI systems³⁵. This process involves evaluating systems against established benchmarks³⁶ or standards to ensure their reliability, safety, and ethical soundness both pre- and post deployment. It encompasses translating identified best practices into specific organization- and system-level

³⁴ 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.

³⁵ 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.

³⁶ 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.

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 monitoring³⁷ – 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.

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 safely 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 good³⁸ 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 responsibly made accessible to third-parties³⁹, 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 deployment and use of AI possible. Investing in hardware makes more computational power available to 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

³⁷ 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.
³⁸ 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.

³⁹ E.g., through responsible or tiered open-sourcing; see Bucknall, B. & Trager, R. (2023). *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements*. Available at: https://cdn.governance.ai/Structured Access for Third-Party Research.pdf

2021 report from the ITU, 2.9 billion people worldwide are without internet access, 96% of them in developing countries. 40 Again, accessibility should be provided in a responsible manner, to avoid the misuse of the resources to create irresponsible AI models and applications.

Building Public Trust in Institutional Oversight of Al

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. 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 trust is publishing a yearly report on the state of AI governance.⁴¹ Such a report would provide a comprehensive overview of advances in AI, highlight challenges, and delineate the steps taken by 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 a year can be a long time in AI, periodic interim reports could complement the yearly report as required to address significant developments that arise between the annual reviews – such as a breakthrough in research, a notable incident, or a change 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 is further crucial in building public trust. While detailed reports are essential for comprehensive understanding, they might be overwhelming for the average citizen. Simplified, clear, and accessible information is needed for everyone to be able to grasp the essence of policy decisions, the rationale behind them, and 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 previously mentioned functions, 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.

⁴⁰ 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/

⁴¹ As proposed by Mulgan et al. (2023) as a function for a GAIO to fulfill, for example.

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, and ensuring benefit to the general public rather than a select few. 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.

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, while being practically relevant and feasible from a technical perspective. While being a soft law function, policy design and implementation assistance paves the way to robust and effective hard laws through deliberation of policy options.

The first mechanism to fulfill this function is the evaluation of (proposed or implemented) 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 while being technically feasible. 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, which in turn would support the expectation setting and harmonization function mentioned previously. This approach would facilitate 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, best practices, tools and guidelines across various sectors and applications. From data privacy to ethical considerations or technical standards, it could offer actionable insights that can be adapted to specific contexts, extending the soft law function 'Provision of Technical Tools Supporting AI Governance'.

Proposal

We view these soft law functions as falling into three baskets: research and information-gathering directed at understanding the impact of the technology; 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 standard-setting bodies such as the IEEE and ISO to provisions of the EU's AI Act⁴² 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.

The initial Framework for the International Governance of AI^{43} developed during a workshop organized by the CCEIA and the IEEE SA, proposed five symbiotic components, addressing the soft law functions outlined above:

- (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.

⁴² 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.

⁴³ _____ (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.

(5.) Creation of a Global AI Observatory (GAIO)⁴⁴, bridging the gap in understanding between scientists and policymakers and fulfilling the functions defined above 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 elucidated 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. Hence, as a first step, this paper proposes to establish a global AI governance mechanism, with immediate efforts focusing on establishing a GAIO, which should initially undertake six areas of activity, as suggested by Mulgan et al. (2023), 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.⁴⁵

To be clear, the structure of this global AI governance mechanism and a roadmap for its implementation must still be developed. Furthermore, developing structures and institutions that can instill a sense of trust, in a landscape dominated by both converging and diverging agendas, will not be easy. Nevertheless, an international body coordinating the activities of institutions fulfilling some of the described functions and taking on the functions that are not currently

_

⁴⁴ 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.

⁴⁵ This could build on efforts such as Stanford's AI Index, see https://aiindex.stanford.edu/report/.

fulfilled by other institutions is a necessary step to ensure the global responsible development and deployment of AI.

Conclusion

The speed at which AI systems 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 to help shape the trajectory of AI development and encourage international cooperation around its normative and technical governance. This paper lists international soft law functions for AI and mechanisms to fulfill them that we need to begin implementing now, presuming they are not being addressed by existing institutions.

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 might serve broader governance needs – including, in the future, for other emerging technologies, including neuroscience and synthetic biology, as well as technologies not yet realized.