IPIE

International Panel on the Information Environment

IPIE Submission to the UN AI Advisory Body

Response to the Interim Report on Governing AI for Humanity

Technical Paper 2024.2

IPIE SUBMISSION TO THE UN AI ADVISORY BODY: RESPONSE TO THE INTERIM REPORT ON GOVERNING AI FOR HUMANITY

Technical Paper TP2024.2

SYNOPSIS

This submission by the International Panel on the Information Environment (IPIE) to the UN AI Advisory Body is in response to a call for submissions regarding its Interim Report on Governing AI for Humanity. The IPIE's Scientific Panel on Global Standards for AI Audits reviewed the report in early 2024 and submitted its recommendations to the Body. The IPIE's response to each section can be found in this report.

Opportunities and Enablers

As the report states, AI offers significant societal benefits if properly managed. It can transform access to knowledge, increase systems' efficiency and human productivity, and contribute to innovation in science, academia, and public service provision.

Without appropriate mechanisms for governance, these opportunities cannot be fully realized. Factors listed as enablers (data, compute, talent) are often concentrated in large, powerful technology companies based in the global north, while the global south, where the global majority lives, is the site of critical and often exploited labor (moderating content, human feedback for reinforcement learning, and other often invisibilized work). These factors make it difficult to assess the real scale of opportunities, reduce the scope for growth by smaller industry players, increases the risk of bias, and heightens concern over a global north and south AI divide. Global oversight of AI is thus key to creating accountability and ensuring that humanity takes full, equal and safe advantage of AI's benefits.

Audits can guide governance:

Audits are crucial to assessing whether these opportunities are being realized and to meaningful accountability, since they operationalize what policy and regulations hope to achieve. Although many actors are developing AI audits, there is no consensus on what should be audited, by whom, or at which state of the system's lifecycle: Government audits generally focus on legal and regulatory compliance, businesses use a risk mitigation framework, and civil society focused on societal good versus harm.

Audits must address both common rails (the enablers needed to develop AI) and guard rails (protection against harm). Some auditors are starting to bridge this divide (for example, the New York Local Law 144 on automated employment decision tools prohibits the use of such tools unless and until they have been satisfactorily audited for bias).

To ensure that opportunities provided by AI can be fully realized a set of agreed parameters must be developed upon which a global audit system can be built. They should include:

1. Fairness versus bias: Are all stakeholders equity represented?

2. Transparency (and auditability): Is access to data open to internal and external stakeholders?
3. Interpretability/explainability: Can the system be understood by users and developers?
4. Regulatory and legal compliance: Does the system comply with all relevant regulations?
5. Data privacy and commons: Ensure a balance between individual privacy and the commons needed for AI to achieve its potential for good.
6. Technical robustness/safety: Is the system secure from attack?

IPIE recommends:

1. A future AI governance regime must account for the value of AI audits to assess the opportunities and benefits presented by them.
2. These audits must be based upon a universally agreed upon set of principles and parameters but must contain adequate flexibility to allow for local diversity.

Risks and Challenges

As the report documents, many risks and challenges are associated with AI. They include mis/disinformation; manipulation; surveillance and invasion of privacy; discrimination; and malign use. To measure these risks, audits are needed. The UK's AI and Digital Hub already makes risk assessment a cornerstone of its AI regulation, as does Singapore's Model AI Governance Framework.

Drawing on Box 3 of the report, we suggest that audits can help assess the risks AI systems pose to:

1. Individuals
    a. Human dignity/value/agency in the form of manipulation, deception, nudging, sentencing
    b. Other human rights and civil liberties, e.g. fair trial (recidivism prediction), presumption of innocence (predictive policing), freedom of expression (nudging), privacy (biometric recognition)
    c. Life opportunities, especially access to education and jobs
2. Groups
    a. Discrimination/unfair treatment of sub-groups, including on basis of gender
3. Society
    a. Information Integrity (mis- or disinformation, deep fakes, personalized news)
    b. Social cohesion (filter bubbles, declining trust in news, information)
4. (Eco)systems
    a. Strain on environment/climate/natural resources
5. Values and norms
    a. Legal norms

Given the scale of potential harm if AI is not properly governed, the development of risk mitigation systems is critical, yet there is no universal agreement around what constitutes 'risk', how to identify it, or how to assess levels. Further complicating the issue is the dearth of data needed to answer these questions due to industry secrecy, technical complexity, and the fragmentation of global governance systems. Well-designed audits will identify, classify and quantify AI vulnerabilities and provide a first step address these concerns. These audits need to be both conducted in-house and/or by second party firms commissioned by them, as well as independent external audits to consider *societal* issues such as commitments to AI responsibility, accountability, or trustworthiness.

IPIE recommends:

1. Audits should identify how AI systems might introduce specific types of risks (as identified above)
2. Audits be conducted by a range of stakeholders including internal and second party auditors, external/independent auditors, community actors, and domain experts. Some governance frameworks are already moving in this direction: the US's Government Accountability Office Framework (2021), for example, lists 'technical stakeholders' (data scientists, program developers, etc), as well as 'a broader community of stakeholders' (policy/legal experts, civil liberty advocates, and end users).
3. AI audits be expansive in aim. The data used to them must include testimonial, physical and documented data, which complement compliance-focused auditing, as well as qualitative data able to capture bias, discrimination and other societal harms.

International Governance of AI

A. Guiding Principles to Guide the Formation of New Global Governance Institutions for AI

The UN Interim Report notes that the complexity of AI systems and governance means there is currently no global consensus around the implementation, interoperability, or incentive systems needed to build and implement a global governance framework. It proposes, as a first step, a series of guiding principles that might be used to inform the creation of a global governance institution. It is crucial that these guiding principles should work in tandem with the established UN's Guiding Principles on Business and Human Rights.

AI audits can support and help operationalize these principles as described below.

Guiding Principle 1: AI should be governed inclusively, by and for the benefit of all

IPIE recommends:

1. A comprehensive AI audit regime be developed to ensure that the opportunities and benefits afforded by AI systems are fairly distributed amongst stakeholder groups. Audits' assessment of stakeholders' relative interaction with and benefit gained from AI can be used to facilitate actions to make 'affirmative and corrective steps…to address the historical and structural exclusion' of these communities (para 47, p 15)

Guiding Principle 3: AI governance should be built in step with data governance and the promotion of data commons

IPIE suggests:

1. The provenance of data used by AI systems must be transparent and must comply with relevant regulations designed to ensure data protection, management, privacy, and indigenous data sovereignty.
2. Audits be used to test whether the '(r)egulatory frameworks and techno-legal arrangements that protect privacy and security of personal data, consistent with applicable laws" (para 51, p 16) are being adhered to by AI systems and their design, development, and use. Such audits must be built upon existing national and multinational systems to take account of existing country/region specific source data and outcome requirements.

Guiding Principle 5: AI governance should be anchored in the UN Charter, International Human Rights Law, and other agreed international commitments such as the Sustainable Development Goals

IPIE suggests:

1. Comprehensive audits that consider issues relating to human rights and the rule of law are conducted to help the UN assess 'AI's impact on a variety of global economic, social, health, security, and cultural conditions' and ensure their alignment with 'universal respect for, and enforcement of, human rights and the rule of law' (para 55, p 15). This can include ensuring that AI systems adhere to principles of transparency, accountability, privacy, and fairness. Conducted in collaboration with affected communities, audits can also help grassroots activists and experts identify the environmental risks of socio-ecological-technological AI systems.

B. Institutional Functions that an International Governance Regime for AI Should Carry Out

In terms of the institutional functions identified in the report, the IPIE Panel on Global Standards for AI Audits and the IPIE more generally, as a global scientific organization that represents over 300 researchers (60% of whom are from the global south and 50% of whom are women), can assist with implementing the following:

Institutional Function 1: Assess regularly the future directions and implications of AI

The report calls for a specialized AI knowledge and research function that, "in a manner similar to the IPCC...would involve an independent, expert-led process that unlocks scientific, evidence-based insights, say every six months, to inform policymakers about the future trajectory of AI development, deployment, and use... This should include arrangements with companies on access to information for the purposes of research and horizon-scanning." The IPIE's Scientific Panel on Global Standards for AI Auditing helps fill this function by building scientific consensus regarding international collaboration on data, compute and talent to help ensure AI systems support the SDGs, international human rights norms, and modern ESG goals. Our reporting and peer review process aggregates knowledge from hundreds of colleagues around the world, in addition to listening calls with experts from civil society, auditing organizations, and corporations. Two of the technical reports we will publish this year--a gap analysis of the current AI auding systems and a report on data provenance related to AI systems and researcher access to that data--speak directly to this function.

Institutional Function 3: Develop and harmonize standards, safety, and risk management frameworks and Function 5: Develop and harmonize standards, safety, and risk management frameworks

The report notes that "several important initiatives to develop technical and normative standards, safety, and risk management frameworks for AI are underway, but there is a lack of global harmonization and alignment." The IPIE brokers international collaboration on data, compute, and scientific expertise; coordinates the reporting and peer review of consensus statements from the scientific community; and elaborates the evidentiary norms, standards, and protocols needed to understand the impact of AI on the global information environment. The IPIE's Scientific Panel on Global Standards for AI Auditing final report will develop international protocols for what an AI audit should be.

Institutional Function 7: Compliance and accountability based on norms

Our comments on the previous sections have explained how audits are crucial to compliance and accountability. We also note that standards, such as the one we are developing in our Panel, can be considered as setting the kind of "non-binding norms" (para 73, p. 18) that might be needed for setting best practices in industry.

Other Comments on the International Governance of AI

The IPIE recommends that audits be considered beyond the transparency of AI models, as the Report describes in the schema for considering interoperability across different AI governance efforts (Figure 1). Audits are crucial to meaningful public accountability in several ways, since they operationalize policy and regulatory aims. They can reveal if the opportunities afforded by AI are being realized, as well as a system's harms. Auditability should be applied to other modules within the four-fold schema, such as Data (to assess adherence to privacy and diversity), Benchmarking (for testing), and Applications (to check for Human Rights due diligence). For example, audits focused on data provenance (using e.g. data sheets and nutritional labels) could evaluate whether the data used respects Indigenous data sovereignty. They could also assess whether the development and deployment of AI systems adheres to the UN's Guiding Principles on Business and Human Rights.

The IPIE also recommends that audits be used to ensure public accountability rather than testing adherence to technical performance criteria or regulatory compliance. This would also involve safe harbor protections for independent audits pressing for such accountability, such as academics and journalists.

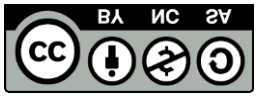
Lastly, the IPIE aims to fulfil the stated need for an IPCC-like body to lead "independent, expert-led process that unlocks scientific, evidence-based insights." As well as producing reports related to global standards for AI auditing, the IPIE aims to identify the biases within existing auditing structures. Experience shows, however, that a critical range of challenges to the SDGs and international human rights norms sit in a broader global information environment. AI and machine learning systems are in a family of emerging and frontier technologies that require applied scientific research about social impact. The IPIE attends to not simply "the future trajectory of AI development, deployment, and use," but other emergent socio-technical concerns.

The IPIE is further able to assert its independence in funding from national governments. This allows the IPIE to function as a neutral and nonpartisan endeavor that can work across privacy regulations, data sharing rules, economic pressures and protectionisms, and the current global political environment. At the same time, the IPIE is crucially global in perspective and involves researchers from around the world. This is especially important in considering the impacts of AI systems, as the data driving AI innovation is often extracted from the global South, the environmental costs of intensive compute are globally borne, and the impact of AI will be on citizens and consumers around the world. The IPIE therefore aims to function as an active steward of the global scientific consensus on matters related to the global information environment, including the impacts of AI systems.

Contributors

Drafting Author: Prem Sylvester (Consulting Scientist, India/Canada). Reviews: Prof. Meredith Broussard (United States), Prof. Wendy Chun (Canada), Prof. Ruben Enikolopov (Russia), Prof. Anikó Hannák (Hungary), Prof. Marcelo Mendoza (Chile), Prof. Alondra Nelson (United States), Prof. Christian Sandvig (United States), Mr. Gbenga Sesan (Nigeria), Mr. Andrew Sporle (New Zealand), Prof. Rohini Srihari (United States), Prof. Christo Wilson (United States), Prof. Mimi Zou (China/United Kingdom).

Copyright Information



Technical Papers

Technical Papers are prepared on specialized topics for which an independent scientific review or technical perspective is deemed essential. Most Technical Papers are used to inform scientific consensus building. Some Technical Papers are prepared as submissions to, or at the request of, international organizations or governments. Technical Papers are produced by individual authors at the request of the IPIE, but any opinions, findings, conclusions, or recommendations expressed in Technical Papers are those of individual authors and may not reflect the views of the international scientific community.

About the IPIE

The International Panel on the Information Environment (IPIE) is an independent and global science organization committed to providing the most actionable scientific knowledge about threats to the world's information environment. Based in Switzerland, the mission of the IPIE is to provide policymakers, industry, and civil society with independent scientific assessments on the global information environment by organizing, evaluating, and elevating research, with the broad aim of improving the global information environment. Hundreds of researchers from around the world contribute to the IPIE's reports. www.IPIE.info.

For more information, please contact the International Panel on the Information Environment (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, P.O. Box, 8034 Zurich, Switzerland.

International Panel on
the Information
Environment

Seefeldstrasse 123
P.O. Box 8034 Zurich
Switzerland