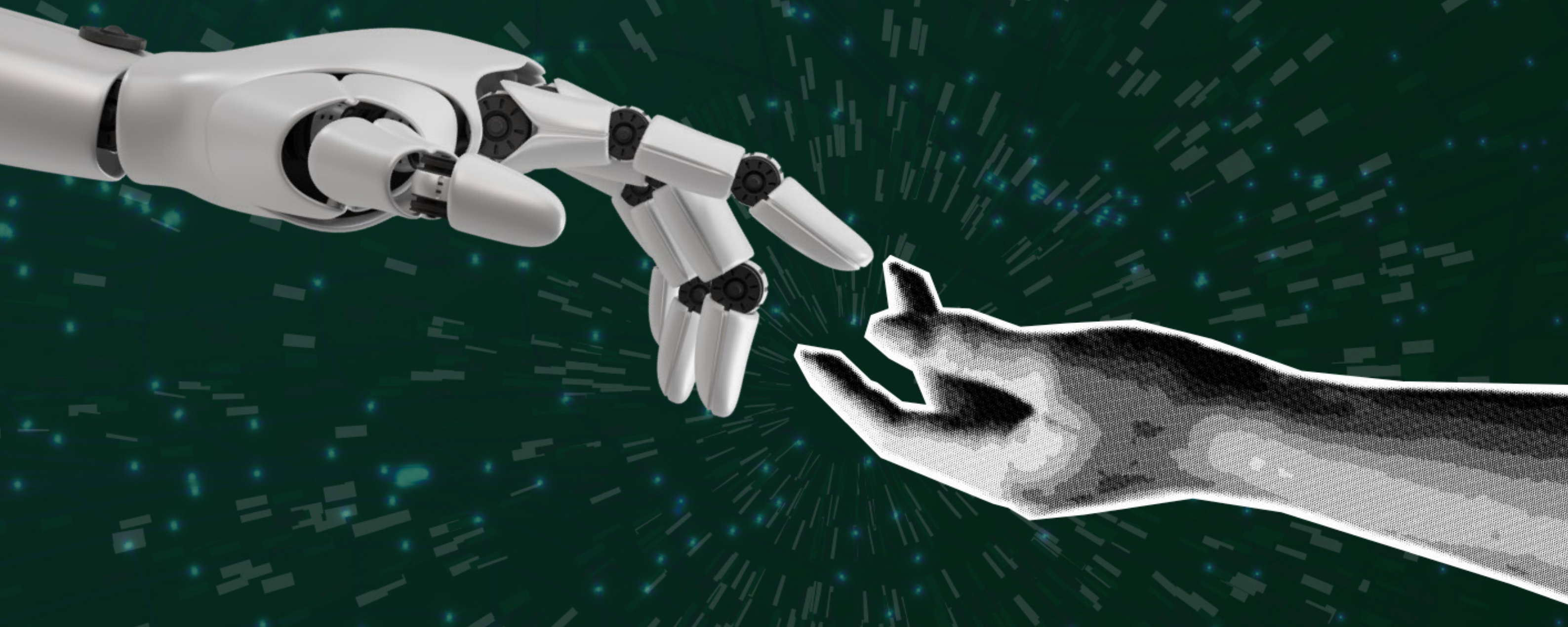
The Dialogue™
INFORM ENGAGE IDEATE

WRITTEN COMMENTS

UN Interim Report on Governing AI for Humanity



MARCH 31, 2024

WRITTEN COMMENTS

UN Interim Report on Governing AI for Humanity

Authors - *Garima Saxena and Jameela Sahiba*

Designer - *Shivam Kulshrestha*

The Dialogue™ is a public policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, on areas around technology and development issues. The Dialogue™ has been ranked as the world's Top 10 think tanks to watch out for, by the Think Tank and Civil Societies Programme (TTCSP), University of Pennsylvania in their 2020 and 2021 rankings.

For more information
https://thedialogue.co

Suggested Citation
Saxena, G & Sahiba, J. (April, 2024) Written Comments: UN Interim Report on Governing AI for Humanity. The Dialogue

Catalogue No
TD/ET/WC/0424/04

Publication Date
April 8, 2024

Disclaimer
The facts and information in this report may be reproduced only after giving due attribution to the authors and The Dialogue™.

Contents

1. Introduction

In response to the United Nations Secretary-General's AI Advisory Body’s invitation for feedback on its Interim Report: Governing AI for Humanity, The Dialogue participated in the consultation process to submit our comments and insights. The submission was made on March 31, 2024. This landmark report underscores the necessity of a globally coordinated approach to AI governance, emphasising the alignment of AI advancements with the principles underpinning international standards and agreements. As The Dialogue continues to monitor the outcomes of this consultation, we remain dedicated to engaging in further discussions and actions that drive the responsible evolution of AI governance. Our engagement in this consultation process reflects our broader mission to bridge divides and foster a more inclusive, equitable, and sustainable future through the responsible development and deployment of AI technologies.

2. Opportunities and Enablers

2.1. Scaling Open-Source AI and Data-Sharing Initiatives

To scale open-source AI models and data-sharing initiatives and make them more accessible to underrepresented communities worldwide, it will be essential to, ***firstly,*** invest in improving internet connectivity, computing resources, and cloud capabilities in underrepresented regions. For the same, it will be critical to partner with telecom service providers, governments, and international organisations to expand digital infrastructure access. Furthermore, open data initiatives that prioritise collecting and curating datasets relevant to underrepresented communities should be supported at the international level. The facilitation of localisation of existing open data and models by involving local stakeholders in translation and annotation processes will also be important.

Secondly, collaborative online platforms connecting underrepresented groups with AI researchers and developers globally could foster peer learning, mentorship programs, and community-driven documentation efforts, enabling effective knowledge sharing. ***Thirdly,*** dedicated funding, such as grants and fellowships, should be allocated specifically for open AI and data projects benefiting underrepresented communities. This funding could provide support to local researchers, developers, and community organisations working in this space. ***Finally,*** facilitating knowledge transfer through exchange programs, fellowships, conferences, and the sharing of best practices and case studies from different regions could accelerate the adoption and adaptation of open-source AI models and data-sharing initiatives.

2.2. Maximising AI's Contribution to the SDGs

AI's contribution to the Sustainable Development Goals (SDGs) can be maximised by ensuring that AI initiatives are rooted in the fundamental principles underlying the SDGs and promoting cross-sectoral collaborations in sectors like energy, climate change, health, and education. Towards the same, dedicated frameworks on AI and SDGs will be indispensable to establish global standards for AI applications and enable scalable solutions adaptable to local contexts. The framework should embed ethical principles and responsible development guidelines to ensure AI advances SDGs without compromising human rights. The framework should also drive essential collaboration between governments, the private sector, and other stakeholders and encourage shared responsibility for ethical and equitable AI development. The framework should also prioritise building individual and community capacity to leverage AI and ensure inclusive education opportunities, reaching underserved communities.

2.3. Addressing Critical Infrastructure Gaps

The critical infrastructure gaps that impact the deployment and development of AI in many regions, especially developing ones, include a lack of reliable high-speed internet, cloud computing resources, affordable access to AI hardware of GPUs (Graphics Processing Units), insufficient energy grids, and advanced cooling systems. Towards the same, providing high-speed internet connectivity, facilitating fibre networks, and partnering with telecoms and organisations that can help expand broadband access will be crucial. Establishing regional cloud computing hubs and data centres and enabling affordable access to AI hardware like GPU will be crucial, and collaborating with manufacturers and exploring innovative financing models like leasing or subscriptions can make hardware more accessible.

Beyond just the digital infrastructure, physical infrastructures like roads, access to water, and electricity will also be essential for setting up and maintaining data centres. There will be a need for skilled personnel not just to manage these technologies but also to repair and maintain them, which poses a challenge in regions lacking educational resources for advanced tech training.

Moreover, establishing educational programs and training initiatives to develop local talent in AI, data science, and related fields will play an essential role in collaboration with universities and technical institutes. Knowledge transfer and exchange programs, where experts from technologically advanced regions can share their expertise and mentor local professionals, should also be encouraged at the international level. Addressing these infrastructural gaps will require international cooperation to harmonise regulations and standards related to AI infrastructure, data privacy, and cybersecurity, enabling cross-border collaboration and interoperability.

3. Risks and Challenges

3.1. Developing Global Standards for AI Transparency

Developing and implementing global standards for AI transparency requires a delicate balance between the need for openness and the protection of proprietary information. One approach is to establish international collaboration among governments, industry leaders, and experts to develop consensus-based standards that ensure transparency while respecting intellectual property rights. These standards should outline clear guidelines for disclosing information about AI systems' functionality, data sources, and decision-making processes without revealing sensitive proprietary details. Additionally, promoting transparency through mechanisms such as standardised documentation and certification processes can enhance trust and facilitate cross-border AI adoption. Standardised documentation ensures that information about AI systems is presented in a consistent and understandable manner, making it easier for users to evaluate their capabilities and limitations. Certification processes, on the other hand, can provide independent verification of AI systems' transparency and adherence to established standards, offering assurance to users and stakeholders about their reliability and trustworthiness.

3.2. Enhancing AI Interpretability and Explainability

Efforts to enhance the interpretability and explainability of AI systems are essential to address their inherent complexity and ensure user trust. This can be achieved through the development of interpretable AI models and techniques like decision trees or LIME explanations that allow users to understand model reasoning that provides insights into how AI algorithms make decisions. Decision trees, for instance, offer a structured, hierarchical representation of decision-making processes, enabling users to trace the logic behind AI-driven outcomes. Similarly, Local Interpretable Model-agnostic Explanations (LIME) provide post-hoc explanations by presenting a set of explanations representing the contribution of each feature to a prediction for a single sample, which is a form of local interpretability.

In addition to technical tools, collaborative research initiatives and interdisciplinary approaches can also contribute to advancing interpretability methods and establishing best practices for explaining AI decisions in understandable terms. Such endeavours fuel the advancement of interpretability methods, driving the evolution of AI systems towards greater transparency and accountability. Interdisciplinary collaborations between AI researchers, cognitive scientists, and domain experts foster a rich exchange of ideas, facilitating the development of best practices for elucidating AI decisions in understandable terms.

Open-source AI models and datasets offer promising avenues for mitigating the challenges posed by proprietary systems by fostering collaboration, knowledge sharing, and innovation. However, there are limitations to open sourcing in competitive commercial environments, including concerns about intellectual property rights, confidentiality, and maintaining a competitive edge. Therefore, while open-source initiatives can promote transparency and accessibility, organisations must carefully evaluate the trade-offs and consider hybrid approaches that balance openness with the protection of proprietary assets.

4. Guiding Principles for International AI Governance Body

4.1. Establishing a Multi-Stakeholder Governance Approach

While much progress has been made towards establishing standards and developing frameworks, most of the existing literature on the risk management of AI focuses on uni-stakeholders, i.e., AI developers. However, given that the adverse implications of AI systems can impact a broader societal level, it is critical to effectively develop, deploy, and operationalise AI systems by taking a systematic approach and considering the AI lifecycle in its entirety. We believe that a principle-based multi-stakeholder approach involving crucial stakeholders in the AI ecosystem will be crucial in balancing innovation and regulation, serving broader public interest across regions and aiding the development of norms on AI ethics and safety. Towards the same, we propose the following approaches:

- A principle-based, multi-stakeholder approach is essential, involving key stakeholders in the AI ecosystem: AI developers, deployers, and impact populations. This approach ensures that AI models are developed to be trustworthy, safe, and fair, prioritise human well-being, and mitigate risks for all stakeholders involved. By acknowledging harms and impacts across the AI lifecycle, we can develop a comprehensive harm and impact map that facilitates the distribution of responsibility and enables appropriate steps to address AI-related challenges.

- The ecosystem approach emphasises the importance of aligning AI efforts and strategies with SDG-related goals, ensuring developments contribute to sustainable progress. This necessitates a government-led, principle-based, multi-stakeholder approach for effective AI regulation, underlining the need for domestic and international coordination. Harmonizing AI principles across sectors is crucial, advocating for consistency and unity at various regulatory levels to ensure AI development and deployment serve the broader public interest across diverse cultural and economic contexts.

- To enforce binding norms on AI ethics and safety across jurisdictions, it will be imperative to establish principles for AI developers and deployers and impact populations at different lifecycle stages, informed by global frameworks. This not only advocates for the harmonisation of AI principles across sectors but also highlights the pivotal role of government in implementing these strategies through enhanced domestic and international coordination.

5. Institutional Principles for International AI Governance Body

5.1. Convening Global Expert Working Groups

As the Interim Report states, the international governance body for AI can take inspiration from the existing global institutions such as IPCC, ITU, CERT and SWIFT. Towards the same, the governance body should, ***firstly,*** convene global expert working groups to regularly assess AI capabilities, risks, and impacts across domains, standardise risk taxonomies, and publish consensus reports synthesising evidence to inform AI governance and policy making. ***Secondly,*** based on the consensus, the body should develop technical AI standards around data formats, define governance standards for ethical AI practices, human oversight, and transparency, and coordinate between standards and regional/national AI regulations for interoperability. ***Thirdly,*** the governance should provide toolkits, best practice guides, and advisory services to upskill AI governance in all countries and facilitate technical cooperation knowledge sharing between leading and emerging AI locales. ***Fourthly,*** the governance body should establish monitoring, incident reporting, and mutual assurance frameworks for high-risk AI systems, coordinate the sharing of high-value AI datasets and models, and compute resources across borders.

The international governance body for AI should comprise experts from diverse fields such as AI and machine learning, ethics, law, sociology, and environmental science, ensuring a comprehensive understanding of AI's impact on different sectors and impact populations. The composition of this body should be rooted in inclusivity, with members representing a broad spectrum of regions, cultures, and economic backgrounds to ensure global representation and perspectives. Furthermore, to tap into a wide range of insights and research findings, the body should facilitate collaborations with universities, research institutions, and private sector entities globally while also engaging with policymakers, industry leaders, and civil society organisations to ensure the body's work is responsive to societal needs and technological advancement.

5.2. Facilitating Technical Cooperation and Knowledge Sharing

Data Sandboxes can be established to create secure environments where researchers can access anonymised or synthetic data for analysis. This will help protect sensitive information while allowing researchers to understand broader trends. International organisations can partner with local organisations to create standardised sandbox protocols for data access and security. Another mechanism that can be explored is federated learning. Through this, we can develop collaborative research models where data remains on company servers, but the training process is distributed. This allows researchers to leverage the collective power of data without compromising confidentiality. Further, companies can be encouraged to release anonymised datasets or open-source some of their AI models for specific research purposes. This can be particularly valuable in areas like healthcare or climate change, where public interest is high. Funding mechanisms to support the development and maintenance of open-source AI tools specifically tailored to the needs of the Global South can also be explored. Lastly, incentive programs can be developed for companies that share data or models for public interest research. This could involve tax breaks, grants, or public recognition. Establish metrics and frameworks to measure the societal impact of responsible data sharing, showcasing the benefits for companies.

thedialogue.co

@_DialogueIndia

@TheDialogue_Official

@The-Dialogue-India

@TheDialogueIndia