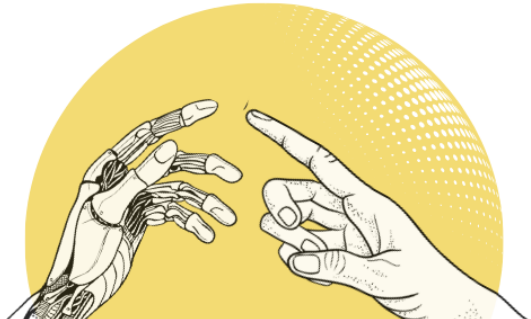
POLICY DIALOGUE

Towards Safe and Trustworthy Artificial Intelligence

A Round Table Discussion

EVENT REPORT

Authors: Mahwash Fatima and Srijan Rai
Published: March 2024



IN THIS REPORT

Executive summary

Against the backdrop of AI's swift integration across various sectors, The Quantum Hub (TQH) convened a roundtable in New Delhi on 21st February 2024, titled '**Towards safe and trustworthy Artificial Intelligence**.'

Deliberating on AI governance and regulation in India, the participants explored challenges and considerations shaping policy frameworks with an emphasis on enabling innovation. Topics included operationalizing international principles of fairness and responsible use in the Indian context, the benefits and risks of legal regulation versus self-regulation, unintended consequences of AI regulation on security and civil liberties, prioritizing implementation efforts, and adapting existing standards. The role of sectoral regulators in ensuring AI compliance was also discussed. With regard to global regulations, the discussion stressed on the importance of learning from diverse jurisdictions and adopting a multi-stakeholder approach.

The conversation on responsible AI underscored the need for AI companies to prioritize local sensitivity and awareness alongside efforts towards red-teaming AI systems for risk identification. Initiatives such as self-regulation frameworks and transparency tools were proposed to foster accountability and transparency. Furthermore, investment priorities in AI within India were discussed, and it was suggested that focus should be placed on enhancing functionality through Learning, Labeling, and Monitoring systems. Discussion on AI development also addressed algorithmic bias and the need to promote diversity in AI development.

The discussion also explored the intersection of AI with copyright and data regulations, highlighting the need for nuanced approaches. Concerns around copyright maximalism and data protection were raised, urging for a balanced approach that fosters innovation while safeguarding individual rights.

Furthermore, the discourse also delved into challenges posed by misinformation, deepfakes, and their implications for electoral integrity. Strategies to raise awareness, particularly in relation to women's online safety and novel approaches to combat misinformation were discussed, emphasizing transparency and accountability in content creation.

The points that emerged from the roundtable discussion acknowledged the numerous opportunities and vast positive potential of AI while underscoring the imperative of responsible AI development, transparency, and collaboration to effectively address its inherent challenges and associated risks.

Discussions on AI governance: Moving towards a trustworthy AI ecosystem in India

India finds itself at the forefront of the global Artificial Intelligence (AI) revolution, poised to harness its transformative potential to drive innovation and economic growth. According to the 'IBM Global AI Adoption Index 2023', 59% of enterprise-scale organizations (over 1,000 employees) surveyed in India are actively using AI in their businesses. According to an EY Study, the adoption of Gen AI could potentially add US$1.2-1.5 trillion to India's GDP over seven years (2023-24 to 2029-30), contributing an additional 0.9% to 1.1% in annual CAGR. However, despite the vast potential of AI, there are significant challenges and risks that require responsible management. While AI presents remarkable opportunities for progress in healthcare, finance, and education, it also comes with inherent risks and complexities. One immediate challenge is the proliferation of misinformation and deepfakes, particularly concerning the upcoming elections. The use of AI by bad actors has the potential to jeopardize the integrity of India's democratic processes and diminish public confidence in them.

Addressing these challenges is not easy, and policymakers confront the formidable task of formulating regulatory frameworks that delicately balance innovation and individual rights. This also requires an assessment of current laws to see if they suffice or if a specialized, nuanced law solely focused on AI is necessary. While existing laws offer some protection, they may not fully accommodate the distinct intricacies of AI, necessitating targeted interventions to bridge potential gaps.

It was in the above context that TQH decided to host a roundtable discussion on 21st February 2024. The event opened with an **insightful keynote address by Member of Parliament, Shri Niranjan Reddy**, who set the tone and context for nuanced discussions on trustworthy AI. This was followed by an engaging discussion between diverse stakeholders, including industry experts, representatives from fact-checking units, civil society organizations, and others on critical questions surrounding AI governance. The key takeaways from these deliberations offer valuable insights that could benefit policymakers and stakeholders alike.

Key takeaways

❖ **Development of safe and responsible AI**

- To address unintended consequences of AI development and deployment, particularly regarding its societal and behavioural implications, it is crucial to focus on red teaming - rigorously challenging the technology to overcome errors, while working in tandem with local teams to understand different contexts. This will also help mitigate bias in AI algorithms.

❖ **AI governance and regulation**

- To be able to leverage the potential of AI, it is essential to strike a balance between facilitating AI uptake and regulating its adoption.
- The need of the hour is thoughtful and measured regulation. AI principles could perhaps be operationalised via a mix of self-regulation and co-regulatory approaches, with guardrails across the value chain. A strong focus will also be required on collaborative efforts to ensure that all relevant stakeholders are included in policy deliberations.

❖ **Mitigating bias in AI**

- **A careful balance is needed between open and closed AI systems/models.** Open models enable possible solutions for designing safe AI, along with AI explainability and transparency in AI outcomes. At the same time, open models could be abused by bad actors. Therefore, developers should assess the benefits and potential risks of open sourcing AI models.

❖ **Addressing the challenge of misinformation and deepfakes**
- When deploying AI systems, it is important to distinguish between exaggerated claims or hyperbole and genuine misrepresentation.

❖ **Global AI regulations and international collaborations**
- While formulating the regulations governing AI, it is important to ensure both harmonization with global standards and contextualisation with the local Indian context. Emphasis must also be laid on evidence-based implementation and capacity building among policymakers to effectively navigate the complexities of AI governance.

DISCUSSION POINTS

Development of safe and responsible AI

The discussion acknowledged the rapid pace of AI development and the potential risks it poses. It was highlighted that there is a strong need for responsible development and deployment of AI technologies. Industry leaders were seen as playing a crucial role in furthering this agenda.

Several suggestions were offered to ensure responsible AI deployment. These included:

- Red teaming: This involves simulating attacks on AI systems to identify weaknesses and vulnerabilities, to help secure the AI models.
- Collaboration with local teams: Working with local teams helps ensure that AI systems are sensitive to specific cultural contexts and potential risks, preventing unintended consequences.
- Self-regulatory frameworks: If the industry could develop clear guidelines, including the formulation of principles on red teaming and benchmarking standards, it could promote transparency and accountability within the development process.

While the discussion acknowledged the possibility of a globally harmonized approach to regulating AI, the overall consensus emphasized the paramount importance of the industry taking responsibility during the AI development process itself.

The discussion also acknowledged the increasing investment in AI by Indian firms, primarily for enhancing functionality through Learning, Labeling, and Monitoring systems (LLMs). There was consensus that these investments should be accompanied with the development of frameworks for ensuring responsible use of the technology.

Mitigating bias in AI

This segment of the discussion centered on the potential harm caused by bias in AI decision-making. It was acknowledged that while AI itself does not inherently harbour bias, its training data can mirror and magnify existing human biases, resulting in unfair and unequal outcomes.

The discussion underscored the challenges in defining and measuring bias, which complicates the development of universal solutions. What proves effective in one context may not necessarily apply well to others. Furthermore, a critical viewpoint cautioned against oversimplifying research by solely

focusing on specific types, such as racial bias. Instead, an approach that recognizes the intricate nature of bias was deemed essential.

A couple of approaches were proposed:

1. Open-source software and weight models: By using open-source resources for training AI models, we can allow others to examine and assess the training resources for potential biases, enabling improvements.

2. Explainable AI (XAI): Implementing XAI which is a methodology/set of processes that provides transparent reasons for outcomes of an AI model, i.e., explaining how an intelligent system viz-a-viz ML model came to a particular decision. However, the discussion acknowledged that XAI has limitations as explanations may be misleading, deceptive, or be exploited by nefarious actors and can also pose privacy risks, as they can be used to infer information about the model or its training data.

Ultimately, the discussion emphasized the need for a multifaceted approach to mitigating bias in AI. This requires incorporating diverse perspectives and methodologies, promoting transparency and accountability in AI decision-making, and continuously developing innovative solutions to address the evolving challenges of bias in AI.

Attribution and copyright issues

Copyright issues emerged as a key concern, exemplified by the music industry's struggles, with companies like Universal Music withdrawing from platforms like TikTok. This highlights the complex landscape of copyright in the context of AI-generated content. The potential for widespread licensing and copyright disputes across diverse industries sparked discussions about finding a balance between protecting creators' rights and encouraging innovation and accessibility. Concerns were raised about overly restrictive copyright regulations, particularly for developing nations, as they could hinder progress.

Text and Data Mining (TDM), which involves using computers to analyze existing digital works for insights, was another point of discussion. Some of the data used for TDM might be copyright-protected, raising the question of whether creators have the right to keep their data outside the scope of this practice. The need for balancing TDM's benefits with respecting creator rights was emphasized. In this context, the discussion acknowledged the global south's perspective, suggesting that developing nations, like India, should advocate for exemptions similar to "fair use" exceptions found in other countries. Fair use exceptions provide a legal justification for using the material that drives a TDM project.

Here, it was pointed out how India currently has a blanket exemption for use of online data which is in contrast to the UK's recent implementation of a tiered system that respects data privacy while also permitting data usage for specific purposes. Such an approach would allow TDM to continue while upholding creators' rights.

The exchange highlighted the need for careful consideration and collaboration among different stakeholders to address the complex challenges and opportunities presented by AI, particularly in

relation to copyright, employment, and protecting individual rights in a rapidly evolving technological landscape.

Addressing the challenge of misinformation and deepfakes

This segment focused on the concerns surrounding deepfakes, particularly their potential to manipulate voters in elections, especially when their AI-generated nature is not evident. The challenges with differentiating between AI-generated content and genuine content can mislead voters and undermine public discourse.

The emphasis on labelling and transparency was seen as a crucial step in mitigating the potential harm that deepfakes can cause, acknowledging, however, that the mere use of AI to create content does not automatically mean misrepresentation.

Another aspect of the discussion delved into the psychological effects of deepfakes, citing the "illusory truth effect" (which refers to the possibility of people believing a content to be true upon repeated exposure to the same) and the risk of widespread exposure to fabricated content over time. While acknowledging the potential of AI in breaking barriers of language and improving access to candidates during elections, participants stressed the need to address misrepresentation and manipulation, particularly when it harms reputations or distorts public discourse.

To combat these risks, the discussion explored several solutions:

- Transparency tools:
    - Visible watermarking: This method helps identify AI-generated content but is subject to limitations like cropping.
    - AI-generated content disclosure tools: These tools can inform users about the use of AI in content creation.
    - Declarations of authenticity: Inspired by existing practices in media, such as real estate advertising that requires declarations of authenticity, it was proposed that this practice be adopted for future content creation.
- Penalties for non-disclosure: To ensure accountability, penalties for not disclosing the use of AI in content creation were discussed.
- Public awareness and education:
    - Strategies to raise awareness about AI, its benefits and risks, particularly on vulnerable groups like women.
    - Emphasizing critical thinking and capacity-building initiatives to equip individuals with the skills to make informed decisions about AI tools.

On the whole, the discussion acknowledged the potential of AI while emphasizing the need for responsible use and robust safeguards against misinformation and misrepresentation in elections and public discourse.

Safe harbour and content moderation

The safe harbour principle, which shields online platforms from liability for user-generated content, sparked a lively debate. Some participants advocated for its continued use with a focus on "*best efforts*"

operation by platforms while arguing against the revision of the said immunity. The discussion acknowledged the formidable challenges of content detection, including the sheer volume of content, difficulty in identifying AI-generated content, and incorrect or misleading results generated by AI models, often referred to as '*AI hallucinations*'.

The conversation then delved into the specific challenges of moderating content generated by AI tools compared to traditional social media content moderation. While traditional moderation typically deals with content disseminated from one-to-many (like social media) or one-to-one (like text messaging), concerns in the AI context grapple with content generated on a one-to-zero level, i.e., content generated by AI tools for the user driven by user instructions/prompts. This distinction suggests potential intervention points at both the input stage (training data) and the output stage (generated content), while acknowledging the non-deterministic nature of AI that can still lead to unpredictable outputs.

There was a shared recognition of the importance of understanding the data used to train generative AI models and the need for explainability in these systems and disclosures about the unreliability of the tools. Overall, the debate highlighted the complexity of revisiting the safe harbour principle and managing AI content moderation, while acknowledging that safe harbour is still existential for the functioning of the internet.

AI governance and regulation

The discussion on AI governance and regulation in India highlighted the challenges and considerations surrounding the development of policy frameworks.

While widely accepted principles like fairness and responsible use of AI are well recognized (e.g., OECD, UNESCO), policymakers struggle to adapt them to India's specific context. While "*operationalization*" can occur through self-regulation or legal regulations, self-regulation is considered problematic due to inconsistent standards across industries. Legal regulations are deemed essential, but nuanced implementation is crucial to avoid overly broad restrictions, as seen in recent times with the IT Act.

The discussion also acknowledged potential unintended consequences, such as the risk of over-regulation while attempting to manage AI-related risks. In the Indian context, adapting existing, well-understood standards and the existing legal framework to address immediate challenges may be more effective than pursuing a universal regulatory framework. For the long term, a predictable and consistent regulatory framework that minimizes fragmentation is essential for effective enforcement. The participants also considered the adequacy of existing sector-specific regulations in addressing AI challenges and emphasized the role of sectoral regulators in ensuring compliance within existing frameworks, potentially minimizing the need for additional measures.

Overall, the discussion emphasized the need for a nuanced and context-sensitive approach to AI governance in India, balancing the benefits of regulation with its potential drawbacks.

Global AI regulations and international collaboration

The discussion on global AI regulation and international collaboration highlighted the diverse approaches in AI governance being adopted across the world, while underscoring the importance of adopting a multi-stakeholder approach.

A participant noted varying regulatory focuses globally, with some regions prioritizing innovation (such as the OECD and UK) and others emphasizing fundamental rights (as seen in the EU). Shifting the focus to India, while some advocated against immediate AI-specific legislation and were of the opinion that regulatory sandboxes, industry standards and self-regulation may be the best way forward, others proposed a co-regulatory framework. This could involve leveraging existing sectoral regulations to enforce standards in specific domains, while concurrently developing comprehensive rules and principles for a broader regulatory framework.

Most importantly, the need for inclusivity in AI discussions was highlighted, stressing the involvement of stakeholders from the global south to ensure comprehensive and equitable governance of AI technologies.

--X--