The Columbia Convening on Openness and AI

moz://a | SIPA | IGP Institute of Global Politics

Policy Readout

27 March, 2024

—

Udbhav Tiwari

Review Contributors:
Kevin Klyman, Madhulika Srikumar, and Stefano Maffulli

Table of Contents

—

On February 29, 2024, Mozilla and the Columbia Institute of Global Politics co-hosted a gathering of over 40 experts and stakeholders in AI to explore the concept of "openness" in the AI era. This diverse group included representatives from leading AI startups, companies, non-profit AI labs, and civil society organizations.

The convening aimed to help develop a better framework for what "open" means in the AI era, drawing inspiration from the pivotal role that open source software has played in technology, cybersecurity, and economic growth over the years. This work is particularly timely: as AI development shifts from research labs into customer-facing products, there has been increased usage of proprietary AI models — raising concerns about negative impacts for innovation, competition, and accountability.

This brief distills the conversations that took place during the workshop for policymakers, identifying benefits that openness can bring to the AI ecosystem, the risk/benefit tradeoffs in opening up different components of the AI stack, and policy recommendations.

Executive Summary

The Convening underscored the significant shift from open source principles to proprietary models in AI development and corresponding concerns over innovation, competition, and accountability. The discussions began by focusing on the many benefits of openness in AI - including enhanced reproducible research, fostering a diverse developer ecosystem and innovation, and promoting inclusion. Additionally, participants also highlighted the role of openness in facilitating accountability, enabling anti-bias research, ensuring security, and improving resource efficiency and competition, among other benefits.

The discussions then moved on to highlight the AI openness ecosystem's complexity, highlighting that AI encompasses more than code, including vast datasets and computing infrastructures. There was consensus around the need to explore and advocate for a multidimensional openness approach - underpinned by the outlook that access to

datasets, model weights, and code benefits developers and researchers by improving data quality, efficiency, and transparency. Participants also acknowledged that bringing openness to different levels of the AI stack brought different levels of risks, which were documented and discussed as a part of discussing responsible release practices and strategic next steps.

Finally on policy recommendations, the gathering stressed the importance of addressing risks and ethical concerns through risk assessments, independent audits, updating legal frameworks for liability, and fostering international collaboration and standards. Many participants also highlighted the need for government intervention and support in nurturing AI research, education, and workforce development and adapting intellectual property laws for open licensing. These recommendations were all ultimately aimed at cultivating a more inclusive and equitable AI ecosystem with agency at its core, echoing the essential role of openness in the advancement of AI technologies for societal benefit.

About the Workshop

During the day-long workshop, participants highlighted the potential of openness in AI to advance societal goals such as ensuring AI safety and effectiveness, fostering innovation and competition, and including underserved communities in the AI ecosystem.

They also emphasised the importance of openness across the entire AI stack, not just within AI models but extending to data, hardware, and user interfaces. The developer community stressed the need for clearer guidance for AI developers on the application of openness principles and called for more nuanced policy discussions on the benefits and risks associated with AI systems that are more open.

Finally, the convening both provided the platform and spotlighted the necessity for a stronger, more organised community committed to promoting, investing in, and advocating for opennesses in AI approaches. A common theme throughout these conversations was how openness can serve as an antidote to market consolidation and safety concerns, especially given the current market dynamics of a few generative AI products dominating the landscape.

Benefits of Openness

Workshop participants agreed that openness in the AI ecosystem catalyses many key benefits., such as:[1]

[1] Risks from openness were also discussed at the workshop and are summarised in the next section.

1. **Enhancing Reproducible Research and Promoting Innovation**

   Openness in AI can help pave the way for reproducible research and shared infrastructure, a cornerstone of scientific advancement. By sharing algorithms, code, and research findings openly, the scientific community can validate and build upon existing work, thereby accelerating the pace of innovation.

2. **Creating an Open Ecosystem of Developers and Makers**

   An AI ecosystem that treats openness as a core principle rather than a risk brings together developers, researchers, and enthusiasts to collaborate on projects, share insights, and solve complex problems. This collective effort accelerates the development of AI solutions and fosters a sense of community among participants, driving innovation through diversity of thought and expertise.

3. **Promoting Inclusion through Open Development Culture and Models**

   An openness-driven AI ecosystem supports the inclusion of diverse datasets and perspectives, which is critical for developing less biassed and representative AI models. This diversity ensures that AI technologies are reflective of the varied human experience and are capable of serving a broad spectrum of society.

4. **Facilitating Accountability and Supporting Bias Research**

   Openness facilitates the auditing of AI systems, making it possible to scrutinise their decisions for fairness, accuracy, and bias. This transparency is essential for holding large companies accountable and for conducting research on bias, environmental impact, and other key challenges.

5. **Fostering Security through Widespread Scrutiny**

   Open systems are subject to widespread scrutiny, which helps in identifying bugs and security vulnerabilities more efficiently than closed systems. This collective vigilance not only improves the security and reliability of AI technologies but also fosters a culture of continuous improvement and iterative innovation.

6. **Reducing Costs and Avoiding Vendor Lock-In**

   Openness in AI can significantly reduce procurement costs for public and private entities by avoiding vendor lock-in. This economic advantage can ensure that AI technologies are accessible to a broader range of organisations, thereby promoting digital autonomy and encouraging the development of local, often privacy preserving, rather than cloud-based architectures.

7. **Equipping Supervisory Authorities with Necessary Tools**

   For regulators and independent researchers, an AI ecosystem underpinned on the principles of openness makes it easier to acquire the tools and access needed to assess large-scale AI systems. This empowerment allows AI technologies to be developed and deployed in a manner that is safe, ethical, and compliant with regulatory standards.

8. **Making Training and Inference More Resource-Efficient, reducing environmental harm**

   Openness in AI also contributes to environmental sustainability by promoting more resource-efficient methods for training and inference. Through shared innovations and collaborative efforts, the AI community can develop techniques that minimise the environmental footprint of AI systems, addressing one of the key challenges in the field.

9. **Ensuring Competition and Dynamism**

   An AI ecosystem with openness at its core can maintain a dynamic and competitive AI sector, providing a level playing field for new entrants and smaller players. This diversity is crucial for meaningfully challenging the dominance of major corporations and for fostering a healthy environment where innovation can thrive without being stifled by monopolistic practices.

10. **Providing Recourse in Decision-Making**

    Lastly, openness in AI ensures that there are mechanisms for recourse when AI-driven decisions negatively impact individuals and communities. In closed systems, affected parties may have no means to challenge or understand decisions made by AI. An open environment allows for decisions to be reviewed, understood, and, if necessary, corrected.

A Multidimensional Approach to the AI Stack

The deliberations at the Columbia Convening highlighted the complexity of the AI openness ecosystem, acknowledging that AI models are not just code but encompass massive datasets, intricate computing infrastructure, and diverse interfaces. This complexity necessitates a multidimensional approach to openness, recognizing the various levels of accessibility and their corresponding benefits. For example, access to pre-training datasets, model weights, and code can empower developers and researchers in different ways, from improving data quality assessments to enabling more efficient model fine-tuning and enhancing transparency in model evaluation.

The following table attempts to crystallise the (non-exhaustive) components of the AI stack that were discussed at the workshop, and the benefits and drawbacks that bringing openness to each of these components would bring to the AI ecosystem.

AI Stack Component	Benefits	Drawbacks
Code • Data pre-processing • Pre-training • Training libraries • Fine-tuning • Inference • Distributed computing • Infrastructure frameworks	• Enhances reproducibility, auditability, and transparency; • Enables better independent assessment of quality and fairness; • Facilitates community participation in model training and understanding training strategies; • Supports open science and knowledge; • Makes inference more efficient and cheaper.	• May expose proprietary technologies or intellectual property; • Could lead to misuse of open-source tools in malicious applications.
Datasets • Pre-training • Supervised fine-tuning • Preference • Evaluation	• Allows examination of biases and fairness; • Promotes understanding of model specialisation and alignment; • Increases transparency in evaluation; • Supports privacy-preserving approaches.	• Risks privacy breaches and security concerns, when the dataset contains personally identifiable information (PII) ; • Potential for harmful content propagation proliferation if datasets are not carefully managed.

Model Weights • Pretrained weights • Intermediate checkpoint weights • Downstream task adaptation • Compressed and adapter • Reward	• Fosters development ecosystems; • Aids in linguistic and cultural model tuning; • Enhances auditability and transparency by enabling independent research and testing; • Supports research in mechanistic interpretability and architecture efficiency.	• Could compromise model integrity if weights are altered maliciously; • Potential for unauthorised use and exploitation of pretrained models.
Documentation • Datasheet • Model cards • Evaluation • Red teaming results • Publications	• Promotes understanding of data representation and model design choices; • Aids in designing efficient models and tracking AI carbon footprint; • Facilitates fair model comparisons; • Increases foresight into potential model misuse.	• Might reveal sensitive or proprietary information; • Could create false sense of trust in a system's capabilities and safety precautions
Distribution • Licence • Type of Release • Acceptable Use Policy/Use Restrictions	• Enables ethical and legal risk management; • Supports staggered scrutiny and societal impact evaluation; • Documents acceptable uses and feedback mechanisms for rights and redress.	• Legal and ethical implications of widespread access; • Strict terms may deter safety research.

Guardrails (applies throughout the stack) • Aligned weights • Programmable weights • Safeguard models/safety classifiers • System prompts	• Improves community testing of safety techniques; • Enables transparent and auditable content moderation; • Facilitates evaluation of content moderation robustness; • Aids in collective determination of effective prompts.	• Risk of enabling adversarial misuse through detailed knowledge of guardrails; • Potential for creation of bypass strategies that undermine safety measures.

Policy Recommendations

Facilitating Openness in AI Development

1. **Include Standardised Definitions of Openness as Part of AI Standards**

   Standardised definitions of openness will be vital for supporting clear and effective regulation, as well as industry best practices to promote appropriate openness and transparency. Governments should facilitate a common understanding among stakeholders, including developers, regulators, and users. By encouraging and supporting community-led standards-development processes, driven by the openness community, to leverage criteria for accessibility, usability, modification rights, and distribution liberties into standards for AI, policymakers can create a more transparent, accessible, and equitable AI landscape.

2. **Promote Agency, Transparency and Accountability**

   Transparency is foundational to building trust and explainability in AI systems. Government AI regulation should incentivize the public disclosure of key information, such as methodologies, training datasets, algorithms, impact assessments, and safety audits. Transparent documentation and change logs help in tracking the development and modifications of AI systems, facilitating easier identification of potential biases or ethical issues.

3. **Facilitate Innovation and Mitigate Monopolistic Practices**

   A competitive ecosystem is essential for fostering innovation and ensuring that the benefits of AI are widely distributed. Openness must be paired with the incentives, resources, and market conditions needed to achieve its full potential.

   Governments should fund R&D for AI projects that practise openness in areas with high societal benefit. They should also initiate active efforts to update antitrust regulations specifically tailored to the digital and AI sectors are crucial to prevent market dominance by a few entities, ensuring a level playing field for new entrants and promoting diversity in AI solutions.

4. **Expand Access to Computational Resources**

   Computational resources are a critical foundation for AI research and development, yet access to these resources is limited by their cost.

   Public-private partnerships can democratise access to cloud computing resources, enabling a wider range of entities to contribute to AI innovation. Investing in supercomputing centres broadens the computational capacity available to researchers and small and medium enterprises, lowering the barriers to entry and fostering a more inclusive research environment. Care should be taken to avoid entrenching the incumbents in the cloud computing market when making these investments.

Addressing Risks and Ethical Concerns

1. **Risk Assessment and Management**

   Early identification and management of risks associated with AI deployments can prevent adverse outcomes and ensure the responsible use of AI technologies.

   Legally mandated risk assessments for certain AI applications that evaluate privacy, security, safety, and human rights implications prior to widespread deployment are essential for understanding and mitigating potential harms. Regulatory bodies, existing domain specific ones or bespoke, should be empowered to review such assessments and ensure compliance.

2. **Independent Audits and Red Teaming**

   Objective evaluation of AI systems through independent audits and red teaming exercises is crucial for identifying vulnerabilities and unintended consequences.

   Governments should spur the development of a standardised framework for conducting such assessments, bringing about consistency and reliability in evaluations. This should be developed with independent inputs beyond large

technology companies. While initially more feasible in domestic settings and more likely to exist in non-binding formats, they should ultimately be coordinated across nations and incorporated into binding law - similar to how the EU's AI Act currently incorporates standards in AI governance.

3. **Privacy and Data Protection**

   It is clear that AI development often involves the processing of vast amounts of personal data, raising significant privacy concerns. Updating privacy legislation to specifically address AI challenges ensures that data protection measures are robust and effective. Legislation should focus on consent, transparency (around data use), and the right to data portability, safeguarding individual privacy while enabling innovation.

4. **Liability and Legal Frameworks**

   The dynamic nature of AI technologies in open source contexts presents challenges in assigning liability when harm occurs and who should be held responsible for such harm.

   An updated legal framework that distinguishes between the responsibilities of different actors involved in AI development and deployment is essential for ensuring accountability. This framework should be flexible enough to adapt to the evolving nature of AI, providing legal certainty that encourages responsible development and deployment.

5. **International Collaboration and Standards**

   Since AI technologies extend past national borders, international collaboration is essential for harmonising standards and regulations.

   Advocating for global cooperation on standards for openness in AI and binding ethical guidelines can facilitate a unified approach to AI governance. Bilateral and multilateral agreements can enhance knowledge exchange and collaboration, promoting the development of globally accepted best practices.

Industrial Policy Interventions

1. **Nurturing AI Research and Development Grounded in Openness**

   Government funding can significantly accelerate the development of AI technologies that are both open and serve the public interest.

   Allocating resources to research institutions and projects with clear societal benefits, similar to other core sciences and the humanities, will foster innovation in areas critical to societal advancement. Encouraging collaborations between

academia and industry leverage can also help bring diverse expertise, fostering cross-sectoral innovation.

2. **Education and Workforce Development**

   Preparing the workforce for the future of AI is essential for harnessing its potential while mitigating risks.

   Investments in education and specialised training programs can equip individuals with the necessary skills to contribute to and responsibly use AI technologies. Promoting STEM education, particularly in AI and machine learning, will facilitate a diverse talent pipeline, crucial for the sustainable development of the AI sector.

3. **Intellectual Property and Open Licensing**

   Intellectual property laws play a significant role in the development, dissemination, and adoption of AI technologies.

   Adapting these laws to support open licensing models can encourage sharing and collaboration in the AI community. This would provide much needed support to the open data community, mitigate legal risk, and facilitate more avenues for independent research into datasets. Guidance on navigating intellectual property challenges will also help ensure that developers can contribute to AI projects without legal uncertainties, promoting innovation and accessibility.

4. **Public Engagement and Stakeholder Involvement**

   Inclusive dialogue on AI development will foster public trust and help ensure that diverse perspectives are considered in policy making.

   Engaging the general public and stakeholders in discussions about the impact of AI will help promote transparency and accountability, rather than the excessive focus currently directed towards the largest AI labs/corporations.. Multi-stakeholder forums can help facilitate collaboration and consensus-building, shaping policies that reflect a broad range of interests and concerns both geographically as well as intellectually.