ENHANCING INTERNATIONAL COLLABORATION TO IMPROVE OPEN-SOURCE AI SAFETY AND SECURITY

PREAMBLE

Artificial intelligence (AI) holds the potential to address some of the world's most pressing challenges and accelerate scientific progress. Open-source AI can facilitate these immense benefits by enhancing access to AI, enabling open collaboration, and fostering inclusive innovation. However, increased access to AI presents distinct risks that require careful consideration by diverse perspectives from around the world, including from governments, academia, industry, and civil society. Moreover, ensuring the safe development and deployment of open-source AI requires global collaboration and cooperation, as no single nation can solve these risks on its own. By working together, countries can build trust and understanding, share knowledge, and create safe open-source AI for the benefit of all, even in a complex global environment.

STATEMENT OF PRIORITIES

COLLABORATION AND INCLUSIVITY

Broaden Participation in Global AI Safety Forums
Global AI safety forums should allow participants from countries without established AI Safety Institutes to contribute, fostering inclusivity and knowledge sharing.

Remove Barriers to Collaboration
Governments, universities, and research labs should strive to promote international collaboration on research on the safety and governance of open-source AI models and avoid imposing restrictions on such research, such as through policy or funding decisions.

Facilitate Global Participation
AI safety conferences should be hosted in accessible locations where researchers from all regions can readily obtain travel visas to reduce barriers to attendance and participation. International scientific reports provide a foundation for understanding risks and rigorous evidence-based decision-making.

Strengthen Engagement Across Nations
Facilitate more inclusive engagement across nations by working with existing international institutions and leveraging neutral platforms.

Collaborate Across Ecosystems and Stakeholders
Foster collaboration between different ecosystems and stakeholders, including academia, government, industry, and civil society, leveraging academia's neutrality.

Promote Continuous Dialogue on the Spectrum of Openness
Promote collaboration between developers involved in open-source and closed-source AI communities to share practices and methods that can facilitate open-source AI safety.

Recognize the Value of Safety and Governance of Open-Source AI
Governments should formally designate the safety and governance of open-source AI as a critical area for international collaboration, highlighting its global significance and their commitment to facilitating this research.

Develop a Shared Vision of AI to Build Trust
Encourage collaboration on areas of AI that build trust, leveraging shared goals.

Close the AI Safety Divide
Ensure underrepresented regions and vulnerable populations have a meaningful voice in AI safety discussions by providing resources and platforms for their participation.

STANDARDS AND OPENNESS

Standardize Incident Reporting

Establish consistent protocols for AI safety incident reporting across countries and institutions to enhance global transparency, accountability, and information sharing for the public good. Consider creating incentives for incident reporting.

Operationalize Standards With Open-Source Tools

Explore the use of open-source tools to promote standards. Define a shared understanding of the scope and risks of open-source AI, ensuring inclusivity across markets.

Clarify Ownership of Standards Development for Open-Source AI

Advance discussions on the role of different ecosystem actors, such as standards development organizations (SDOs) and AI Safety Institutes (AISIs), in defining standards for open-source AI.

Address Cultural and Contextual Nuances

Address cultural and national nuances related to risk, such as different conceptions of terms including safety and security.

AI SAFETY PROCESSES

Create an Enduring, Inclusive AI Safety Process

Consolidate discrete AI safety initiatives and events (e.g., academic conferences, government summits, and open letters) into an enduring, inclusive, community-driven process that allows for flexibility to account for different perspectives. Ensure consistency and scalability in AI safety processes across nations. This process should not be owned by any one country.

Prioritize Critical Risks

Prioritize high-impact, high-likelihood risks and promote international consensus on addressing adversarial risks.

Develop Reliable Testing and Evaluation

Create robust, reliable methods for testing and evaluating open-source AI models and systems that can be trusted globally, even between countries with strained relationships.

Engage Model Developers Globally

Encourage AI model developers to participate in safety discussions with culturally, geographically, and politically diverse stakeholders to enhance mutual understanding.

Address Marginal Risks From Open-Source AI

Researchers should focus on the marginal increased risks associated with the broad accessibility and availability of open-source AI, as informed by international scientific reports, ensuring these risks are effectively mitigated to prevent widespread harm.

ABOUT THIS STATEMENT

This statement reflects ideas developed by a group of international experts at a workshop held in December 2024. The workshop used the **Chatham House Rule**, therefore neither the identities nor the affiliations of the participants may be revealed or confirmed without their permission. In addition, individuals who participated in the workshop may not agree with some or all parts of the statement.