CONCORDIA AI
安远AI

info@concordia-ai.com
www.concordia-ai.com

AI Action Summit: Public Interest AI

1. What are different approaches to AI auditing and how should these be implemented in order to be taken up at scale? What building blocks may be missing to make the market for AI auditing (e.g. certification schemes, training, standardization)? What types of other auditing parallels in history might be taken as examples?

To advance international progress on AI safety auditing, the AI Action Summit should:

1. Create a 'Global AI Safety and Trust Fund' to improve the scientific rigor of AI safety evaluations.
2. Launch international technical exchanges between evaluation institutions, both governmental and non-governmental.
3. Create an international working group to advance the consensus on AI safety and trust testing beyond the AI Action Summit.

(1) The Summit should broker new institutional partnerships between universities and research centers worldwide and establish a 'Global AI Safety and Trust Fund' to support international academic collaboration.

These projects would address a critical gap — AI safety constitutes only about 2% of current AI research despite growing international calls for enhanced collaboration on safety, ethics, and societal impact.[1] New joint research centers and personnel exchanges between universities and/or research institutes would improve expertise around the world. A global fund in the range of US$100 million would be the largest such fund for AI safety and trust, which could support ambitious research efforts on AI alignment, evaluation, ethics, and more. These initiatives would substantially advance global cooperation without requiring binding government agreements.[2]

[1] Emerging Technology Observatory on AI safety. For more on the need for global academic cooperation, see The Manhattan Declaration on Inclusive Global Scientific Understanding of AI, which was co-chaired by Turing Award Winner Yoshua Bengio and former White House OSTP Acting Director Alondra Nelson, and signed by Concordia AI CEO Brian Tse.

[2] One example of such institutional partnerships is the Global Partnership on AI's Expert Support Centers. Similarly, the UN High-Level Advisory Body (HLAB) on AI has proposed a 'Global Fund for AI' that would include

info@concordia-ai.com
www.concordia-ai.com

(2) Hold a workshop at the Summit focused on AI trust and safety evaluations, involving national AI safety institutes and relevant government-backed AI evaluators around the world.

While it may be premature to establish formal international standards on AI testing given the technology's rapid evolution, building common understanding and interoperability is crucial. The Summit should promote an "agile governance" approach that allows standards to evolve with our understanding of AI risks and opportunities. This could begin with countries experimenting domestically while maintaining international dialogue channels to begin building interoperability around AI trustworthiness and safety testing. A workshop at the Summit could seek technical consensus on questions such as what types of evaluation methodology (e.g. static benchmarks, human uplift) are most appropriate for evaluating different types of risks and what risk levels should be considered acceptable. AI safety evaluations bodies around the world could also hold a joint AI testing demonstration at the Summit.

Chinese participation is essential to these efforts. Chinese organizations have demonstrated significant expertise in evaluating AI risks across domains from bias and privacy to autonomous cyberattacks and dual-use AI in science.[3] Their work aligns with Chinese government statements indicating support for dialogue on AI safety, particularly regarding national security and public safety concerns. Chinese policymakers have specifically highlighted the risk of terrorist misuse of AI, emphasized the importance of human control over AI systems, and called for creating an AI safety oversight system.[4]

(3) Establish an international working group to further define acceptable risk thresholds and harmonize global evaluation methods after the Summit.

---

safety and governance funding, see Governing AI for Humanity. The fund could include contributions from stakeholders such as national governments, technology companies, and philanthropists. A US$100 million fund would be the largest such fund for AI safety and trust, compared for instance to the Frontier Model Forum's US$10 million+ AI Safety Fund. Similar efforts to improve pandemic resilience through a World Bank "Pandemic Fund" have already raised US$2 billion.

[3] See Concordia AI's China's AI Safety Evaluations Ecosystem. Additional papers on benchmarking dual-use AI in science that were published after this piece include SciSafeEval and SciKnowEval.

[4] See Concordia AI's State of AI Safety in China pages 29-32 and What does the Chinese leadership mean by "instituting oversight systems to ensure the safety of AI?".

info@concordia-ai.com
www.concordia-ai.com

An international working group, created at the Summit, could take the lead on improving AI testing interoperability and report results on common safety standards at the subsequent global AI summits. Without additional international dialogue and progress on testing standards, there will be no way to ensure that these testing mechanisms meet a common quality level.

2. The past decade of work on AI has shown that the term "public interest" encompasses many aspects, including accountability, social justice, human rights, consumer protection, antitrust tools, refusal and curtailment of applications, environmental justice, audits, redressal mechanisms, among others. What existing definitions of 'public interest AI' would enable the broader field to come to shared values, goals and outcomes?

The AI Action Summit should advance three key initiatives on public interest AI:

1. Publish an action plan and list of pilot projects to advance AI for Sustainable Development Goals (SDGs).
2. Announce pilot projects for AI safety and trust as a global public good.
3. Commission a global, authoritative survey of public views on AI around the world.

(1) Develop an action plan for applying AI to meet the SDGs at the Summit and announce a list of pilot projects over the following 6-12 months.

The global SDGs have been adopted by all UN member states, but the world is "woefully off track" in reaching these goals.[5] Concordia AI joined leading experts in calling for AI to play a pivotal role in achieving SDGs in domains such as health and education.[6] The Summit's action plan should move beyond analysis to concrete implementation.

(2) Affirm AI safety and trust as a global public good in the Summit joint statement and launch demonstration projects.

[5] UN, Halfway to 2030, world 'nowhere near' reaching Global Goals, UN warns.

[6] See The Manhattan Declaration on Inclusive Global Scientific Understanding of AI, which was co-chaired by Turing Award Winner Yoshua Bengio and former White House OSTP Acting Director Alondra Nelson.

info@concordia-ai.com
www.concordia-ai.com

AI safety and trust should be treated as a global public good, as it is non-rivalrous (use by one nation does not diminish availability for others) and non-exclusive (benefits inherently cross borders). This framework suggests three models for international collaboration:

- Aggregate efforts: shared incident reporting and red teaming.
- Weakest link initiatives: universal safety guardrails.
- Single best-shot opportunities: breakthrough safety research and evaluation methods.

The Summit should announce support for this concept in its joint statement and launch specific trial projects for each model within 12 months - for example, establishing a shared incident reporting system, developing common safety standards, and creating joint research initiatives for breakthrough safety methods.

(3) Commission a comprehensive global survey on AI development and trustworthiness, particularly engaging Global South perspectives.

While some studies exist, including Concordia AI's analysis of Chinese AI surveys, there is insufficient understanding of how global publics view AI development and safety.[7] Moreover, insights are heavily skewed towards Western populations. The Summit should commission a comparative and truly global survey, leveraging expertise from organizations like Missions Publiques (France) and the Center for International Security and Strategy (China). Results should be delivered within 12 months to inform future governance decisions.

3. What are existing efforts to define what openness in AI means, led by whom, bringing together which stakeholders and across which regions?

On AI openness, the AI Action Summit should:

1. Recognize the complex and non-binary nature of the open vs. closed debate.
2. Establish a framework that grants qualified scientific researchers expanded access to advanced AI models.
3. Require third-party auditing of more risky AI models and publication of system cards.

[7] See Concordia AI's State of AI Safety in China pages 68-73.

info@concordia-ai.com
www.concordia-ai.com

4. Promote dissemination of open-source AI safety and trust toolkits.

(1) The Summit's joint statement should acknowledge nuances in the open vs. closed source debate.

Concordia AI believes that AI openness exists on a spectrum, from fully closed systems to completely open models with published weights, code, and training data.[8] While openness promotes innovation and prevents power concentration among private actors, it requires careful management to mitigate potential misuse. Effective policymaking on openness requires acknowledging these complexities.

(2) The Summit should gain developer agreement for controlled access by the scientific community to AI models while maintaining security measures.

The Summit could develop standardized protocols for granting vetted scientists and auditors access to advanced AI models. The system could be managed by an international committee of scientists and industry actors. Such access would facilitate academic research, peer review, and analysis of risks. The Summit should also call for greater public research on the risks and benefits of openness for the most advanced AI models.[9]

(3) The AI Action Summit should establish mandatory transparency requirements and independent oversight mechanisms to prevent concentration of power among private AI companies.

These transparency requirements should include third-party auditing, detailed model card disclosures, and wider adoption of government model registries like China's registration system for generative AI.[10] The Summit could fund independent evaluations through a 'Global AI Safety and Trust Fund' and create an international database for sharing AI model registry information across jurisdictions. Given the

[8] Concordia AI discussed these topics in a Chinese-language report titled Responsible Open-Sourcing of Foundation Models, written in collaboration with the Peking University Institute for Artificial Intelligence and the Beijing Institute of General Artificial Intelligence.
[9] Similarly, the UN HLAB on AI report Governing AI for Humanity called for examination of the "limits (if any) of open-source approaches to the most advanced forms of AI."
[10] See Concordia AI State of AI Safety in China report pages 10-13.

info@concordia-ai.com
www.concordia-ai.com

general-purpose nature of these systems, transparency requirements should be tied to model capabilities in addition to specific high-risk use cases. The Summit could also announce a commitment by companies or states to support organizations conducting independent, third-party ratings of corporate frontier AI commitments.[11] Through these measures, the Summit can ensure AI developers uphold their safety and security commitments while serving the public interest.

(4) The Summit should establish a centralized repository of open-source AI safety tools.

Safety technology must be widely accessible to be effective. This initiative would provide funding and support for building new tools or improving existing tools, such as Shanghai AI Lab's CompassKit, UK AI Safety Institute's Inspect toolkit, Singapore AI Verify Foundation's Project Moonshot, and the COMPL-AI Framework for the EU AI Act.[12] Through these coordinated efforts, the Summit can help establish widely-usable global tools for responsible AI development.

6. Do you have any other comments or suggestions on Public Interest AI?

Our understanding of what constitutes "public interest" should take into account the interests of future generations.[13] We recognize that the decisions, actions, and inactions of present generations regarding AI development and governance have an intergenerational multiplier effect. Therefore, we should ensure that present generations act with responsibility towards safeguarding the needs and interests of future generations.

In the climate domain, this could take the form of creating a Carbon-Efficient AI Model Leaderboard and encouraging smaller scale AI models at the Summit. The recent trend towards aiming to develop ever-larger and more powerful general-purpose AI models will likely place greater demands on energy usage and harm the environment.[14] To promote greener AI development, the AI Action Summit could build a Carbon-Efficient AI Model Leaderboard to encourage a race-to-the-top

[11] For instance, see the Centre for the Governance of AI's A Grading Rubric for AI Safety Frameworks and SaferAI's Risk Management Maturity of AI companies.
[12] CompassKit; Inspect; Project Moonshot; COMPL-AI Framework.
[13] See the UN Summit of the Future's Declaration on Future Generations.
[14] International Scientific Report on the Safety of Advanced AI: Interim Report.

info@concordia-ai.com
www.concordia-ai.com

dynamic among developers.[15] Chinese developers are already innovating in this area – DeepSeek V2 is an open-source model with inference costs one-seventh of Llama 3 70B and one-seventeenth of GPT-4 Turbo, while ModelBest's MiniCPM-Llama3-V 2.5 is only 8B parameters yet performs well on multimodal capabilities and credibility.[16]

[15] See pilot projects such as the AI Carbon Efficiency Observatory and Hugging Face on Energy Scores for AI Models.
[16] DeepSeek; Synced Review.