CONCORDIA AI
安远AI

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

AI Action Summit: AI of Trust - Security & Safety

1. How can we improve the scientific understanding of AI risks and opportunities?

To strengthen scientific assessment of AI, the AI Action Summit should:

1. Mandate inclusion of social science and humanities expertise alongside technical experts in major AI assessment and evaluations efforts.
2. Require long-term scenario planning (around 5+ years) in AI risk assessments, particularly for impacts on future generations.

(1) The Summit should declare in the joint statement the importance of including diverse and socio-technical perspectives in AI evaluation.

Current AI evaluations often focus narrowly on technical benchmarks while overlooking critical real-world impacts. For example, automation's effects vary significantly across industries and regions, requiring expertise beyond technical metrics. Additionally, existing evaluations inadequately address multilingual and multicultural issues. The Summit should ensure that future iterations of the International Scientific Report on the Safety of Advanced AI and the potential Independent International Scientific Panel on AI in the UN integrate social sciences and humanities expertise to fully understand these complexities.

(2) The Summit should commission exercises to develop future AI scenarios and ensure that such analysis is included in potential AI evaluations efforts under the UN or global AI summits.

Any assessment should incorporate a long-term, future-generations perspective, but current evaluations, like the AI Risk Global Pulse Check for the UN AI Advisory Body, typically look only one to two years ahead.[1] There is uncertainty around the appropriate time horizon for AI forecasting, and longer-term projections confront increased uncertainty. Nevertheless, the current short-term focus is insufficient

[1] UN High-Level Advisory Body on AI, Governing AI for Humanity.

given AI's potential multi-generational effects – from climate impact of large models to the potential for developing vastly more powerful systems. Drawing lessons from the International Panel on Climate Change's scenario planning, the Summit should commission analyses of AI scenarios beginning at least several years in the future.[2]

2. What specific technical standards and benchmarks for AI trustworthiness, explainability, and fairness should be developed and implemented globally? How can we ensure these standards are adaptable to rapid AI advancements?

To advance international AI evaluation standards, the AI Action Summit should:

1. Launch regular dialogues between testing institutions worldwide, including Chinese organizations, to build consensus on evaluation methodologies and risk thresholds.
2. Create a working group to develop an international guide for AI content watermarking.

(1) 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

[2] See IPCC and Carbon Brief.

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]

(2) Begin developing a guide for AI content provenance.

Content provenance and watermarking represent a concrete opportunity for international cooperation. A key Chinese standards body has already released implementation guidelines requiring both visible and invisible watermarks for AI-generated content, with further regulations in development.[5] This approach could help inform creation of international standards on watermarking, which are needed given the global span of many content providers and platforms. The Summit can establish a working group of technical experts who would develop an AI watermarking guide to inform companies globally.

3. How can we define and implement clear, enforceable thresholds for AI development and deployment, considering both current and potential future capabilities? What mechanisms should be in place to regularly review and update these boundaries?

To improve risk thresholds around catastrophic AI risks, the AI Action Summit should pursue three pillars of work:

1. Broker agreement on international red lines and early warning indicators regarding AI misuse and loss of control, such as autonomous replication, weapons development, and cyberattacks.
2. Promote continuous AI safety testing for warning indicators by creating an international working group on standards and hosting a joint evaluation exercise among diverse countries.
3. Kick off discussions around developing a set of crisis management protocols that can be triggered if certain risk thresholds are crossed.

[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?".

[5] See Concordia AI's AI Safety in China #2 and AI Safety in China #17.

(1) The Summit should declare a set of risk thresholds focusing on transnational, catastrophic risks.

This targeted approach to AI restrictions has gained substantial international support. The UN AI Advisory Body has called for investigating "thresholds for tracking and reporting of AI incidents," specifically highlighting bans on uncontrollable systems, untraceable AI, and potential "superintelligence."[6] Leading AI experts from China and the West have similarly endorsed establishing red lines for the most dangerous capabilities.[7]

(2) Create an international working group and hold a demonstration evaluation to promote AI safety testing for early warning indicators.

Continuous AI safety testing for early warning indicators is needed to identify when systems approach risk thresholds and ensure that red lines are not crossed. To ensure that these testing mechanisms meet a common quality level, AI safety evaluations bodies around the world should hold a joint AI testing demonstration at the Summit. An international working group, created at the Summit, could take the lead on improving AI testing interoperability and ensuring common safety standards in the following year and report results at the subsequent global AI summits.

(3) The final pillar is developing a set of crisis management protocols that can be triggered if certain risk thresholds are crossed.

The world will need contingency plans in the event that risk thresholds are crossed. The global AI summits should begin holding discussions on this topic as soon as possible, so that global consensus has been achieved by the time risks become extreme. The AI Action Summit should set up a working group to explore this topic more deeply in future summits. Potential contingency plans could include mandating further AI safety research, assurances, and human oversight until proven safe. There should be special attention and support for Global South countries in building resilience to risks.[8]

[6] UN High Level Advisory Body for AI, Governing AI for Humanity.
[7] International Dialogues on AI Safety, IDAIS-Beijing, 2024 Statement.
[8] For additional context, see Concordia AI at the AI Seoul Summit.

4. What concrete mechanisms and reporting standards should be established to ensure AI companies transparently demonstrate adherence to global AI ethics commitments they made at previous Summits?

To strengthen the transparency of frontier AI development, the AI Action Summit should:

1. Publicly track company adherence to voluntary commitments on a website.
2. Transform voluntary corporate safety commitments into mandatory requirements through coordinated government regulation.
3. Establish a 'Global AI Safety and Trust Fund' to support independent evaluation networks and third-party ratings.
4. Begin discussions on creating a global AI governance body to ensure consistent oversight across jurisdictions.

(1) Create a public website to track company implementation of commitments.

Such a measure would strengthen the requirement in the Seoul Frontier AI Safety Commitments for companies to publish safety frameworks at the AI Action Summit.[9] While the 2024 Seoul Summit's Frontier AI Safety Commitments were a positive step, relying solely on corporate self-regulation is insufficient for protecting public safety and national security.[10] The Summit should also seek commitments among countries to implement minimum transparency standards from AI companies domestically and share some of this information in an international database.

(2) Coordinate domestic government regulation to strengthen corporate voluntary commitments.

Governments could mandate greater transparency and testing of models of a certain size or risk level in domestic legislation. For instance, China already possesses an AI model registry and security review system, and Chinese experts have proposed various ideas to enhance transparency into AI models and

[9] Frontier AI Safety Commitments, AI Seoul Summit 2024.

[10] See Concordia AI at UK Global AI Safety Summit.

companies.[11] The AI Action Summit could hold a discussion among governments about the barriers to mandatory disclosure and testing requirements for advanced AI models.

(3) Build a robust third-party evaluation capacity through establishing a 'Global AI Safety and Trust Fund.'

The Summit could also announce a US$100 million 'Global AI Safety and Trust Fund' to support international academic collaboration, with a significant focus on evaluations.[12] While some domains like cybersecurity and CBRN-E (Chemical, Biological, Radiological, Nuclear, and Explosive) have established expertise, global expertise on emerging risks like deceptive alignment and autonomous replication is insufficient. The 'Global AI Safety and Trust Fund' would support developing evaluation networks across both established and nascent domains. It could also promote the creation of third-party organizations to rate frontier AI commitments for ensuring independent accountability of AI companies.[13]

(4) Establish a dialogue mechanism for creating a global AI governance body.

The France AI Commission's report argues for creating a World AI Organization that "would share scientific findings on the workings and effects of AI, and define binding standards for AI systems and how they should be audited."[14] We agree that international coordination is becoming increasingly critical as companies plan to scale frontier AI systems 100-1000x (in terms of effective compute) in the next three to five years.[15] Without coordinated licensing and oversight, countries risk a regulatory race to the bottom, as seen with global corporate taxes. As these systems' potential for misuse and loss of control

[11] See Concordia AI's State of AI Safety in China pages 11-13 and State of AI Safety in China Spring 2024 Report pages 57-58.

[12] Similarly, the UN High-Level Advisory Body (HLAB) on AI has proposed a 'Global Fund for AI' that would include safety and governance funding in Governing AI for Humanity. The fund could entail 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 "Pandemic Fund" have already raised US$2 billion.

[13] 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.

[14] France Artificial Intelligence Commission, Our AI: Our Ambition for France.

[15] Epoch AI, Training Compute of Frontier AI Models Grows by 4-5x per Year.

grows, we must establish global governance mechanisms before private companies' decisions irreversibly impact humanity's future.

5. How can we foster effective international cooperation in AI governance, including the creation of a global network of AI evaluation and safety institutes? What should be the key objectives, structure, and funding mechanisms for such a network?

The AI Action Summit should communicate that an inclusive international network of AI safety institutes (AISIs) is needed that:

1. Develops common standards and a body of knowledge on AI evaluations, with participation from all major AI powers, including China.
2. Creates an early warning system for indicators of dangerous model capabilities.

(1) The Summit should achieve agreement that the AISI network will include an appropriate level of participation from all major AI powers, balancing security concerns with safety benefits.

The foundation for this network already exists through AISIs' shared focus on model evaluation. These institutes serve three core functions: foundational safety research, domestic standards-setting, and international coordination.[16] Building on this base, cooperation can occur at varying levels of depth depending on national security sensitivity – from sharing evaluation science to developing best practices, creating common testing datasets, and conducting joint projects.

While some deep coordination (like sharing model weights or bioweapon testing datasets) may raise national security concerns, most evaluation cooperation has limited sensitivity. Historical precedent shows that competitor countries can still cooperate on safety technology - for example, the US shared nuclear safety mechanisms with the Soviet Union during the Cold War.[17] Similarly, sharing AI evaluation insights and best practices would help all countries avoid accidentally developing unsafe systems.

[16] See Institute for AI Policy and Strategy, Understanding the First Wave of AI Safety Institutes: Characteristics, Functions, and Challenges.

[17] Jeffrey Ding, Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies.

China's participation is crucial for effective global oversight. Several Chinese institutions already perform AI safety functions, including government-affiliated think tanks and labs that evaluate cybersecurity risks, model autonomy, and scientific dual-use risks.[18] Excluding major AI powers from better safety practices and scientific understanding would fundamentally undermine global AI safety.

(2) Obtain a commitment in the Summit declaration that AISI network members will share early indicators of dangerous model capabilities with other network members.

Safety evaluations by AISIs will provide early warning indicators when AI systems near catastrophically unsafe levels. The AISIs should develop a system to share information about such indicators when they are tripped, similar to early warning systems for pandemics, so that key global policymakers know when we may be approaching red lines anywhere in the world. This network could include national AI safety institutes and other organizations working on safety testing from a public good perspective. In China, while there is not yet an official AI safety institute, several institutions have comparable functions; for instance, at least one major government-affiliated think tank and two major government-funded AI labs conduct testing on threats such as cybersecurity and jailbreaking.[19]

[18] See Concordia AI's China's AI Safety Evaluations Ecosystem. Examples of scientific dual-use evaluations include Control Risk for Potential Misuse of Artificial Intelligence in Science by researchers from Microsoft Research Asia, University of Science and Technology of China, and Nanyang Technological University, as well as Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science by researchers from Yale University, Shanghai Jiao Tong University, National Institutes of Health, Mila-Quebec AI Institute, and ETH Zurich.

[19] A non-exhaustive list of relevant evaluation efforts include: government-backed think tank China Academy of Information Communications Technology's AI safety benchmark, state-backed Shanghai AI Lab's SALAD-Bench benchmark, and the Beijing Academy of AI's FlagEval.