Department for
Science, Innovation
& Technology

Guidance

Frontier AI Safety Commitments, AI Seoul Summit 2024

Updated 7 February 2025

This publication is available at https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

The UK and Republic of Korea governments announced that the following organisations have agreed to the Frontier AI Safety Commitments:

- Amazon
- Anthropic
- Cohere
- Google
- G42
- IBM
- Inflection AI
- Meta
- Microsoft
- Mistral AI
- Naver
- OpenAI
- Samsung Electronics
- Technology Innovation Institute
- xAI
- Zhipu.ai

> The following organisations have been added to the existing list:

- Magic
- Minimax
- 01.ai
- NVIDIA

The above organisations, in furtherance of safe and trustworthy AI, undertake to develop and deploy their frontier AI models and systems[footnote 1] responsibly, in accordance with the following voluntary commitments, and to demonstrate how they have achieved this by

publishing a safety framework focused on severe risks by the upcoming AI Summit in France.

Given the evolving state of the science in this area, the undersigned organisations' approaches (as detailed in paragraphs I-VIII) to meeting Outcomes 1, 2 and 3 may evolve in the future. In such instances, organisations will provide transparency on this, including their reasons, through public updates.

The above organisations also affirm their commitment to implement current best practices related to frontier AI safety, including: internal and external red-teaming of frontier AI models and systems for severe and novel threats; to work toward information sharing; to invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights; to incentivize third-party discovery and reporting of issues and vulnerabilities; to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated; to publicly report model or system capabilities, limitations, and domains of appropriate and inappropriate use; to prioritize research on societal risks posed by frontier AI models and systems; and to develop and deploy frontier AI models and systems to help address the world's greatest challenges.

Outcome 1. Organisations effectively identify, assess and manage risks when developing and deploying their frontier AI models and systems. They will:

I. Assess the risks posed by their frontier models or systems across the AI lifecycle, including before deploying that model or system, and, as appropriate, before and during training. Risk assessments should consider model capabilities and the context in which they are developed and deployed, as well as the efficacy of implemented mitigations to reduce the risks associated with their foreseeable use and misuse. They should also consider results from internal and external evaluations as appropriate, such as by independent third-party evaluators, their home governments[footnote 2], and other bodies their governments deem appropriate.

II. Set out thresholds[footnote 3] at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable. Assess whether these thresholds have been breached, including monitoring how close a model or system is to such a breach. These thresholds should be defined with input from trusted actors, including organisations' respective home governments as appropriate. They should align with relevant international agreements to which their home governments are party. They should also be accompanied by an explanation of how thresholds were decided upon, and by specific examples of situations where the models or systems would pose intolerable risk.

III. Articulate how risk mitigations will be identified and implemented to keep risks within defined thresholds, including safety and security-related risk

mitigations such as modifying system behaviours and implementing robust security controls for unreleased model weights.

IV. Set out explicit processes they intend to follow if their model or system poses risks that meet or exceed the pre-defined thresholds. This includes processes to further develop and deploy their systems and models only if they assess that residual risks would stay below the thresholds. In the extreme, organisations commit not to develop or deploy a model or system at all, if mitigations cannot be applied to keep risks below the thresholds.

V. Continually invest in advancing their ability to implement commitments i-iv, including risk assessment and identification, thresholds definition, and mitigation effectiveness. This should include processes to assess and monitor the adequacy of mitigations, and identify additional mitigations as needed to ensure risks remain below the pre-defined thresholds. They will contribute to and take into account emerging best practice, international standards, and science on AI risk identification, assessment, and mitigation.

Outcome 2. Organisations are accountable for safely developing and deploying their frontier AI models and systems. They will:

VI. Adhere to the commitments outlined in I-V, including by developing and continuously reviewing internal accountability and governance frameworks and assigning roles, responsibilities and sufficient resources to do so.

Outcome 3. Organisations' approaches to frontier AI safety are appropriately transparent to external actors, including governments. They will:

VII. Provide public transparency on the implementation of the above (I-VI), except insofar as doing so would increase risk or divulge sensitive commercial information to a degree disproportionate to the societal benefit. They should still share more detailed information which cannot be shared publicly with trusted actors, including their respective home governments or appointed body, as appropriate.

VIII. Explain how, if at all, external actors, such as governments, civil society, academics, and the public are involved in the process of assessing the risks of their AI models and systems, the adequacy of their safety framework (as described under I-VI), and their adherence to that framework.

---

1. We define 'frontier AI' as highly capable general-purpose AI models or systems that can perform a wide variety of tasks and match or exceed the capabilities present in the most advanced models. References to AI models or systems in these commitments pertain to frontier AI models or systems only.

2. We define “home governments” as the government of the country in which the organisation is headquartered.
3. Thresholds can be defined using model capabilities, estimates of risk, implemented safeguards, deployment contexts and/or other relevant risk factors. It should be possible to assess whether thresholds have been breached.