

16 AI companies sign Frontier AI Safety Commitments at the AI Seoul Summit

Source copy kept by wheresthe.ai for the story "16 AI companies sign Frontier AI Safety Commitments at the AI Seoul Summit" (<https://wheresthe.ai/d/intl-seoul-frontier-ai-safety-commitments-2024-05/>).

Source: <https://gov.uk/government/news/historic-first-as-companies-spanning-north-america-asia-europe-and-middle-east-agree-safety-commitments-on-development-of-ai>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 086a142d2f8ba1acbe4dedb0b75c8295d02c695eff26e0ebe8c4aaa924f80492

Press release

Historic first as companies spanning North America, Asia, Europe and Middle East agree safety commitments on development of AI

New commitments to develop AI safely have been agreed with 16 AI tech companies spanning the globe.

From:

Department for Science, Innovation and Technology , The Rt Hon Michelle Donelan and The Rt Hon Rishi Sunak MP

Published: 21 May 2024

2022 to 2024 Sunak Conservative government ">

This was published under the 2022 to 2024 Sunak Conservative government

Global AI companies agree safety commitments.

The UK and Republic of Korea have secured commitment from 16 global AI tech companies to a set of safety outcomes, building on Bletchley agreements with expanded list of signatories.

in the extreme, leading AI tech companies including from China and the UAE have committed to not develop or deploy AI models if the risks cannot be sufficiently mitigated

agreement also commits companies to ensuring accountable governance structures and public transparency on their approaches to frontier AI safety

New commitments to develop AI safely have been agreed with 16 AI tech companies spanning the globe, including companies from the US, China and the Middle East, marking a world-first on the opening day of the AI Seoul Summit (Tuesday 21 May).

As two days of talks get underway, Zhipu.ai (China) and the Technology Innovation Institute (UAE) are among companies that have signed up to the fresh 'Frontier AI Safety Commitments':

Amazon

Anthropic

Cohere

Google / Google DeepMind

G42

IBM

Inflection AI

Meta

Microsoft

Mistral AI

Naver

OpenAI

Samsung Electronics

Technology Innovation Institute

xAI

Zhipu.ai

Where they have not done so already, AI tech companies will each publish safety frameworks on how they will measure risks of their frontier AI models, such as examining the risk of misuse of technology by bad actors.

The frameworks will also outline when severe risks, unless adequately mitigated, would be “deemed intolerable” and what companies will do to ensure thresholds are not surpassed.

In the most extreme circumstances, the companies have also committed to “not develop or deploy a model or system at all” if mitigations cannot keep risks below the thresholds.

On defining these thresholds, companies will take input from trusted actors including home governments as appropriate, before being released ahead of the AI Action Summit in France in early 2025.

The 16 companies who have agreed to these commitments represent the most significant AI technology companies around the world, including representation from the US and China, the world’s two biggest AI powers.

Prime Minister, Rishi Sunak, said:

It’s a world first to have so many leading AI companies from so many different parts of the globe all agreeing to the same commitments on AI safety.

These commitments ensure the world’s leading AI companies will provide transparency and accountability on their plans to develop safe AI .

It sets a precedent for global standards on AI safety that will unlock the benefits of this transformative technology.

The UK’s Bletchley summit was a great success and together with the Republic of Korea we are continuing that success by delivering concrete progress at the AI Seoul Summit.

Technology Secretary Michelle Donelan said:

The true potential of AI will only be unleashed if we’re able to grip the risks. It is on all of us to make sure AI is developed safely and today’s agreement means we now have bolstered commitments from AI companies and better representation across the globe.

The UK is a world leader when it comes to AI safety, and I am continuing to galvanise other nations as we place it firmly on the global agenda and capitalise on the Bletchley Effect.

With more powerful AI models coming online, and more safety testing set to happen around the world, we are leading the charge to manage AI risks so we can seize its transformative potential for economic growth.

Republic of Korea Minister Lee said:

Ensuring AI safety is crucial for sustaining recent remarkable advancements in AI technology, including generative AI , and for maximizing AI opportunities and benefits, but this cannot be achieved by the efforts of a single country or company alone.

In this regard, we warmly welcome the ‘Frontier AI Safety Commitments’ established by global AI companies in collaboration with the governments of the Republic of Korea and the UK during the ‘ AI Seoul Summit’, and we expect companies to implement effective safety measures throughout the entire AI lifecycle of design, development, deployment and use.

We are confident that the ‘Frontier AI Safety Commitments’ will establish itself as a best practice in the global AI industry ecosystem, and we hope that companies will continue dialogues with governments, academia, and civil society, and build cooperative networks with the ‘ AI Safety Institute’ in the future.

These commitments build on the groundbreaking agreements made with leading AI tech companies at Bletchley Park during the first AI Safety Summit six months ago, as well as other existing commitments such as the US Voluntary Commitments and the Hiroshima Code of Conduct.

Tino Cuellar, President of the Carnegie Endowment for International Peace said:

As the world continues to wrestle with opportunities and risks from frontier AI models, governments, private actors, and civil society all have key roles to play and must find productive ways to work together. Efforts like the safety commitments announced at the Seoul Summit will play a central role in strengthening effective governance and helping countries strike a sensible balance between innovation and safety.

Tom Lue, General Counsel and Head of Governance, Google DeepMind said:

These commitments will help establish important best practices on frontier AI safety among leading developers. The agreement demonstrates the value of focused international Safety Summits, where scientifically-grounded conversations can take place.

Reid Hoffman, co-founder of LinkedIn and Inflection. AI , said:

AI is and will make massive improvements to human life and work. But it is also very important to navigate the risks. That's why these commitments are such a crucial step forward in managing the most severe risks of advanced AI .”

We applaud the UK and Republic of Korea for ensuring that developers globally implement the state of the art in frontier AI safety. We look forward to discussing our safety framework alongside other companies at the upcoming France AI summit.

Peng Zhang, CEO of Zhipu.ai said:

Artificial General Intelligence (AGI) holds the promise of transforming numerous aspects of our lives, but with this advanced technology comes the crucial responsibility of ensuring AI safety. As we delve deeper into the realms of AGI , it is imperative that we prioritize the development of robust safety measures to align AI systems with human values and ethical standards, thereby safeguarding our future in an AI -driven world.

Professor Yoshua Bengio, World leading AI researcher, Turing Prize winner, Lead author of the International Scientific Report on the Safety of Advanced AI , said

I am pleased to see leading AI companies from around the world sign up to the Frontier AI Safety commitments. In particular, I welcome companies' commitments to halt their models where they present extreme risks until they can make them safe as well as the steps they are taking to boost transparency around their risk management practices.

This voluntary commitment will obviously have to be accompanied by other regulatory measures, but it nonetheless marks an important step forward in establishing an international governance regime to promote AI safety.

Ben Garfinkel, Director, Centre for the Governance of AI said:

These commitments represent a crucial and historic step forward for international AI governance. My expectation is that they will speed up the creation of shared standards for responsible AI development, help the public to judge whether individual companies are doing enough for safety, and support informed policy making around the world.

David Zapolsky, Senior Vice President of Global Public Policy and General Counsel, Amazon said:

Amazon is proud to endorse the Frontier AI Safety Commitments, which in many ways represent the culmination of a multi-year effort to establish global norms for the safe, secure, and trustworthy development and deployment of frontier AI . As the state of the art of AI continues to evolve, we agree that it is important for companies to provide transparency about how they are managing potential risks of frontier models and honoring their global commitments.

Ya-Qin Zhang, Chair Professor and Dean, Institute for AI Industry Research, Tsinghua University

I strongly welcome the world's leading AI companies committing to managing the most severe risks posed by frontier models. These commitments by a diverse group of Chinese, American and international firms represent

significant step forward on the public transparency of AI risk management and safety processes.

Gillian Hadfield, Schwartz Reisman Chair in Technology and Society at the University of Toronto said:

While the capabilities of emerging models are rapidly evolving, it is clear that the public and government leaders lack visibility to be able to assess and mitigate the risks posed by frontier AI models. The Frontier AI Safety Commitments represent a significant step towards tangible and effective regulation of AI, demonstrating a joint commitment to best practices in AI safety, increasing public transparency, and offering flexibility to allow for change as the landscape evolves.

Anna Makanju, VP of Global Affairs, OpenAI said:

The Frontier AI Safety Commitments represent an important step toward promoting broader implementation of safety practices for advanced AI systems, like the Preparedness Framework OpenAI adopted last year.

The field of AI safety is quickly evolving and we are particularly glad to endorse the commitments' emphasis on refining approaches alongside the science. We remain committed to collaborating with other research labs, companies, and governments to ensure AI is safe and benefits all of humanity.

Chris Meserole, Executive Director, Frontier Model Forum said:

The commitments announced today are a significant step forward for frontier AI safety – proactively identifying, assessing and managing risks is essential to the safe development and deployment of the most capable AI systems. We look forward to working with industry, government, and the scientific community to turn the commitments into practice.

Dr. Najwa Aaraj, CEO, Technology Innovation Institute said:

The age of AI has arrived and the opportunities for society are immense. The power of generative AI and large language models is already transforming industries, but for us to reap the maximum benefit, we must keep trustworthiness and safety at the core of the technology's development. The Technology Innovation Institute is a firm believer of trust and secure AI, committed to open source its large language models, and I am delighted to join the other global AI players here in Seoul to discuss and outline the roadmap for AI and set the direction for a safe but prosperous future for us all.

Dan Hendrycks, Safety Advisor to xAI said:

These voluntary commitments establish that the major AI companies around the globe agree on basic safety standards. This helps lay the foundation for concrete domestic regulation.

Professor Yi Zeng, Director of Center for Long-term AI, China, said:

These commitments should not only be welcomed in principle, but also be supported for actions. Assessing risks across the full lifecycle of AI and setting out risk thresholds for meaningful, effective and sufficient human control are the cores for raising the levels of safety for Frontier AI.

Experiences for actionable risk assessment need to be shared broadly so that mistakes do not repeat again and again across different institutes, companies, and countries. Thresholds need to be interoperable so that we are weaving a web of safe AIs, ensuring the safety not only for the self, but also for others, and for all of humankind.

Brad Smith, Vice Chair and President, Microsoft said:

In 2016, Microsoft began the work to implement a principled and human-centered approach to advancing AI systems in a safe, secure, and trustworthy manner. The Frontier AI Safety Commitments are an important acknowledgement of how safety frameworks must help to address risks that may emerge at the frontier of AI development, especially as their capabilities advance. The tech industry must continue to adapt policies and practices, as well as frameworks, to keep pace with science and societal expectations.

Christina Montgomery, Chief Privacy and Trust Officer, IBM:

IBM believes that effective regulation coupled with corporate accountability will allow businesses and society at large to reap the benefits of AI. As such we are proud to continue our international engagement and our commitment to safe and responsible development of these technologies via the AI Seoul Summit.

Brian Tse, Founder and CEO, Concordia AI, Beijing-based social enterprise focused on AI safety and

governance, said:

The Frontier AI Safety Commitments represent a crucial step forward in the risk management of advanced AI models. Building on the foundation laid by the Bletchley Declaration, the Commitments hold frontier AI developers accountable for the risks posed by their most powerful systems. I look forward to working with AI developers, governments, third-party evaluators, and other stakeholders to ensure the highest standards of AI safety are upheld for the benefit of humanity.

Beth Barnes, founder and head of research at METR, globally leading research non-profit for Frontier AI model safety, said:

We think it's vital to get international agreement on the "red lines" where AI development would become unacceptably dangerous to public safety without adequate mitigation. We're excited to see many parties agreeing to set out such red lines in the Frontier AI Safety Commitments. We admire the UK and South Korea's leadership in establishing these commitments.

Michael Sellitto, Head of Global Affairs at Anthropic said:

The Frontier AI safety commitments underscore the importance of safe and responsible frontier model development. As a safety-focused organization, we have made it a priority to implement rigorous policies, conduct extensive red teaming, and collaborate with external experts to make sure our models are safe. These commitments are an important step forward in encouraging responsible AI development and deployment.

Nick Clegg, President, Global Affairs at Meta said:

Ensuring that safety and innovation go hand and hand is more critical than ever as industry makes massive strides in developing AI technology. To that end, since Bletchley last year, we've launched our latest state-of-the-art open source model, Llama 3, as well as new open-source safety tooling to ensure developers using our models have what they need to deploy them safely. As we've long said, democratizing access to this technology is essential to both advance innovation and deliver value to the most people possible. Ahead of next year's Summit, we look forward to continued streamlining of international initiatives to ensure a global approach to responsible AI .

Aidan Gomez, Co-founder and CEO of Cohere said:

We are grateful for the UK and Republic of Korea's leadership developing a framework to address potential risks associated with frontier AI models. Cohere is encouraged that in the months since Bletchley Park, the UK, and the industry generally, have increasingly focused on the most pressing concerns, including mis- and dis- information, data security, bias and keeping humans in the loop. It is essential that we continue to consider all possible risks, while prioritizing our efforts on those most likely to create problems if not properly addressed.

Kiril Evtimov, Group CTO, G42, said:

G42 is proud to join this coalition of companies dedicated to advancing AI safely and responsibly. By committing to rigorous safety frameworks and transparent governance, we are not only safeguarding our technological advancements but also paving the way for a future where AI benefits all of humanity. This agreement underscores our collective responsibility and the power of international collaboration in shaping the ethical development of AI .

Notes to editors

The 'Frontier AI Safety Commitments' can be found in full here .

More information on previous commitments made at Bletchley Park's AI Safety Summit can be found here .

On 21 and 22 May, the United Kingdom and the Republic of Korea are hosting the AI Seoul Summit. It will bring together international governments and select global industry, academia and civil society leaders for discussions across two days.

It builds on the inaugural AI Safety Summit hosted by the United Kingdom at Bletchley Park in November last year and will be one the largest ever gatherings of nations, companies and civil society on AI .

On day one, President Yoon Suk Yeol of the Republic of Korea and Prime Minister Rishi Sunak are co-chairing a virtual session for world leaders on innovation and inclusivity, as well as commitments made at Bletchley. On day two, Minister of Science and ICT , H.E. Lee Jong Ho of the Republic of Korea, and the Secretary of State for

Science, Innovation and Technology, Michelle Donelan, will co-chair a ministers' session with representatives from countries, the European Union and the UN, alongside key figures from industry, academia and civil society looking at AI safety, sustainability and resilience.

Share this page

The following links open in a new tab

Share on Facebook

(opens in new tab)

Share on X

(opens in new tab)