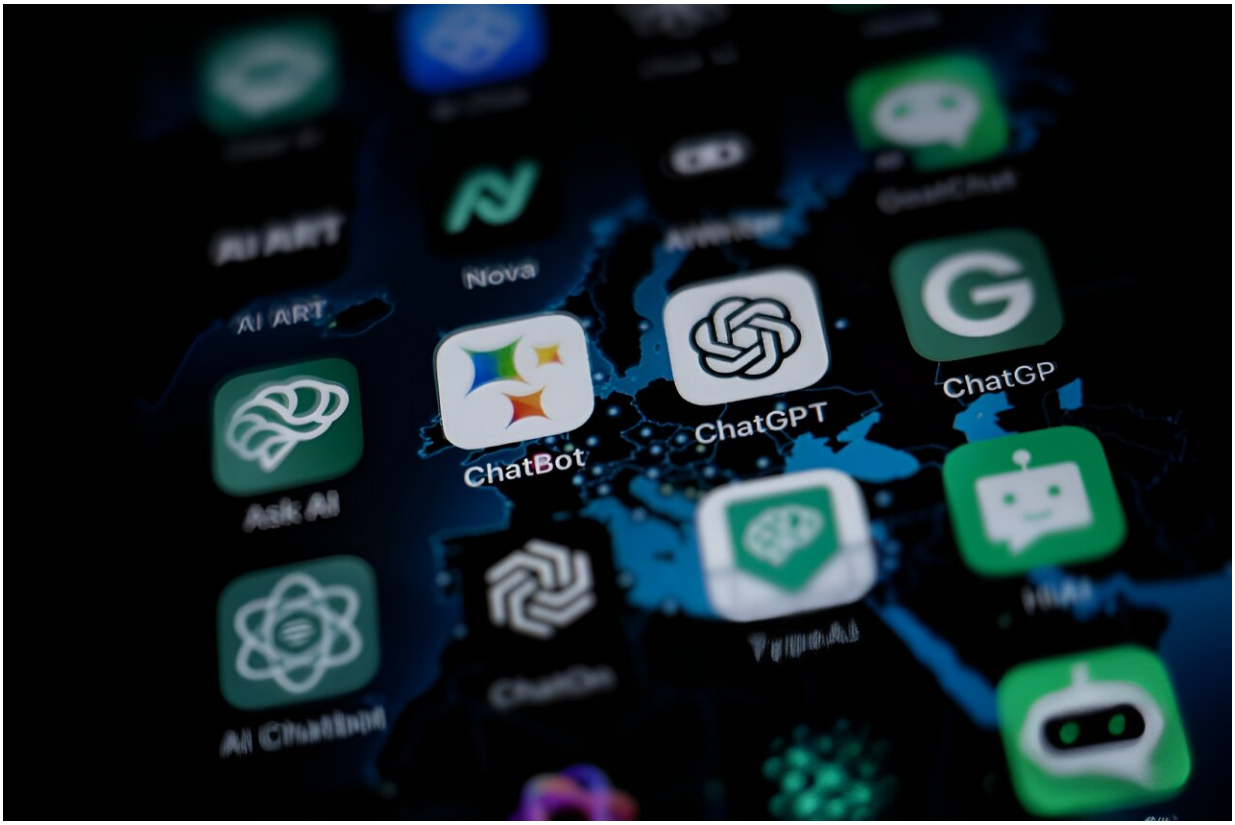
16 top AI firms make new safety commitments at Seoul summit

May 21 2024, by Qasim NAUMAN



More than a dozen of the world's leading AI firms have made fresh safety commitments at a global summit in South Korea.

More than a dozen of the world's leading artificial intelligence firms made fresh safety commitments at a global summit in Seoul on Tuesday,

the British government said in a statement.

The agreement with 16 tech firms—which include ChatGPT-maker OpenAI, Google DeepMind and Anthropic—builds on the consensus reached at the inaugural global AI safety summit at Bletchley Park in Britain last year.

"These commitments ensure the world's leading AI companies will provide transparency and accountability on their plans to develop safe AI," UK Prime Minister Rishi Sunak said in a statement released by Britain's Department for Science, Innovation and Technology.

Under the agreement, the AI firms that have not already shared how they assess the risks of their technology will publish those frameworks, according to the statement.

These will include what risks are "deemed intolerable" and what the firms will do to ensure that these thresholds are not crossed.

"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," the statement added.

The definition of these thresholds will be decided ahead of the next AI summit, due to be hosted by France in 2025.

The firms that have agreed on the safety rules also include US tech titans Microsoft, Amazon, IBM and Instagram parent Meta; France's Mistral AI; and Zhipu.ai from China.

In his opening remarks, South Korea's President Yoon Suk Yeol flagged "growing concerns over potential risks and negative impacts of AI, including fake news through deepfake and the digital divide."

"Since the digital space is hyper-connected and transcends national borders, we need digital norms at the global level," he added.

Danger of 'deepfakes'

The stratospheric success of ChatGPT soon after its 2022 release sparked a gold rush in generative AI, with tech firms around the world pouring billions of dollars into developing their own models.

Generative AI models can generate text, photos, audio and even video from simple prompts, and its proponents have heralded them as a breakthrough that will improve lives and businesses around the world.

But critics, rights activists and governments have warned that they can be misused in a wide variety of situations, including the manipulation of voters through fake news stories or so-called "deepfake" pictures and videos of politicians.

Many have called for international standards to govern the development and use of AI, and have called for action at summits such as the two-day gathering in Seoul this week.

In addition to safety, the Seoul summit will discuss how governments can help spur innovation, including into AI research at universities.

Participants will also consider ways to ensure the technology is open to all and can aid in tackling issues such as climate change and poverty.

The Seoul summit comes days after OpenAI confirmed that it had disbanded a team devoted to mitigating the long-term dangers of advanced AI.

"The field of AI safety is quickly evolving and we are particularly glad

to endorse the commitments' emphasis on refining approaches alongside the science," Anna Makanju, OpenAI's vice president of global affairs, said in the statement announcing the new commitments on Tuesday.

The two-day summit will be partly virtual, with a mix of closed-door sessions and some open to the public in Seoul.