

Meta, Google, and OpenAI reveal their safety plans following White House summit

Seven of AI's top companies attended a White House summit Friday, reaching a deal with the Biden administration to roll out new guardrails to enhance the safety and security of their systems and users, like watermarking AI-generated content.

"These commitments are real, and they are concrete," President Joe Biden at the White House Friday following the meeting. "Artificial intelligence is going to change the lives of people around the world. The people here will be critical for shepherding that innovation with responsibility and safety by design."

Earlier in the day, the seven companies – including Amazon, Anthropic, Meta, Google, Inflection, and OpenAI – made voluntary, sometimes vague, commitments to invest in research and safety, security stress-testing, and assist in third-party audits of system vulnerabilities.

"These commitments are real, and they are concrete."

"AI should benefit the whole of society. For that to happen, these powerful new technologies need to be built and deployed responsibly," Nick Clegg, Meta's president of global affairs, said in a statement Friday. "As we develop new AI models, tech companies should be transparent about how their systems work and collaborate closely across industry, government, academia and civil society."

Most of the companies included in the announcement issued

statements confirming their commitments while declining to provide explicit details into how they planned to employ all of them. Currently, the White House has not provided any enforcement mechanisms for ensuring companies live up to the promises.

Anthropic was the least descriptive, saying in a statement Friday, "We all need to join in a race for AI safety. In the coming weeks Anthropic will share more specific plans concerning cybersecurity, red teaming, and responsible scaling, and we hope others will move forward swiftly as well."

One of the companies' most consumer-facing commitments is a new White House request to place watermarks on AI-generated content. In a call with reporters on Thursday, a White House official said that they expect these watermarks to apply to audio and visual content created by individual users.

In a Friday blog post, OpenAI said that the watermarking agreements would require the companies to "develop tools or APIs to determine if a particular piece of content was created with their system." It also said that previously easily audiovisual content, like AI voice assistants, would not be covered by this commitment.

Google pledged to deploy similar disclosures . Kent Walker, Google's president of global affairs, reiterated that commitment in a statement Friday, saying that the company would "soon be integrating watermarking, metadata, and other innovative techniques into" upcoming generative systems.

Many of the companies included in the White House's announcement have already made similar promises to safeguard their tech prior to Friday's summit. On Tuesday, Meta said that it would open-source its large language model Llama 2, making it free for researchers and most commercial use, similar to OpenAI's GPT-4, which powers ChatGPT and Microsoft

Bing. Meta also agreed to open its model for public evaluation at this year's Def Con.

"This is a serious responsibility, and we've got to get it right," Biden said at the White House Friday.

<https://www.theverge.com/2023/7/21/23803244/meta-google-openai-microsoft-artificial-intelligence-ai-white-house-commitments>