

US AI Safety Institute agrees pre-release model testing with OpenAI and Anthropic

Source copy kept by wheresthe.ai for the story "US AI Safety Institute agrees pre-release model testing with OpenAI and Anthropic" (<https://wheresthe.ai/d/us-aisi-openai-anthropic-testing-agreements-2024-08/>).

Source: <https://nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research>

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

SHA-256 of the saved text: 6ac49661f0e279dba986efa492ef4b19433c332afe7d1a20017bb363c7126471

U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI | NIST

[Skip to main content](#)

Official websites use .gov

A .gov website belongs to an official government organization in the United States.

Secure .gov websites use HTTPS

A lock (

) or <https://> means you've safely connected to the .gov website. Share sensitive information only on official, secure websites.

<https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research>

NEWS

U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI

These first-of-their-kind agreements between the U.S. government and industry will help advance safe and trustworthy AI innovation for all.

August 29, 2024

Share

Facebook

Linkedin

X.com

Email

In June 2025, AISI was re-established as CAISI by Secretary of Commerce Howard Lutnick.

GAITHERSBURG, Md. — Today, the U.S. Artificial Intelligence Safety Institute at the U.S. Department of Commerce's National Institute of Standards and Technology (NIST) announced agreements that enable formal collaboration on AI safety research, testing and evaluation with both Anthropic and OpenAI. Each company's Memorandum of Understanding establishes the framework for the U.S. AI Safety Institute to receive access to major new models from each company prior to and following their public release. The agreements will enable collaborative research on how to evaluate capabilities and safety risks, as well as methods to mitigate those risks. "Safety is essential to fueling breakthrough technological innovation. With these agreements in place, we look forward to beginning our technical collaborations with Anthropic and OpenAI to advance the science of AI safety," said Elizabeth Kelly, director of the U.S. AI Safety Institute. "These agreements are just the start, but they are an important milestone as we work to help responsibly steward the future of AI." Additionally, the U.S. AI Safety Institute plans to provide feedback to Anthropic and OpenAI on potential safety improvements to their models, in close collaboration with its partners at the U.K. AI Safety Institute. The U.S. AI Safety Institute builds on NIST's

more than 120-year legacy of advancing measurement science, technology, standards and related tools. Evaluations under these agreements will further NIST's work on AI by facilitating deep collaboration and exploratory research on advanced AI systems across a range of risk areas. Evaluations conducted pursuant to these agreements will help advance the safe, secure and trustworthy development and use of AI by building on the Biden-Harris administration's Executive Order on AI and the voluntary commitments made to the administration by leading AI model developers. About the U.S. AI Safety Institute The U.S. AI Safety Institute , located within the Department of Commerce at the National Institute of Standards and Technology (NIST), was established following the Biden-Harris administration's 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence to advance the science of AI safety and address the risks posed by advanced AI systems. It is tasked with developing the testing, evaluations and guidelines that will help accelerate safe AI innovation here in the United States and around the world.

Super intelligence and Information technology

Released August 29, 2024, Updated May 4, 2026

Was this page helpful?