

International Network of AI Safety Institutes holds its launch meeting in San Francisco

Source copy kept by wheresthe.ai for the story "International Network of AI Safety Institutes holds its launch meeting in San Francisco" (<https://wheresthe.ai/d/intl-ai-safety-institutes-network-launch-2024-11/>).

Source: <https://www.nist.gov/node/1867351>

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

SHA-256 of the saved text: 055b97854e9b5bf8f103c45a4129f16eb6073cdd2f8512255cfd64d32b5e8dcb

FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco | 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/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international>

NEWS

FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco

November 20, 2024

Share

Facebook

LinkedIn

X.com

Email

The Network will drive alignment on and build the scientific basis for safe, secure, and trustworthy AI innovation around the world.

Ahead of the convening, the Network is announcing key developments including a joint mission statement, more than \$11 million in funding toward synthetic content research, findings from the Network's first multilateral testing exercise, and a joint statement on risk assessments of advanced AI systems.

This technical-level working meeting gathers network members and industry, academic, and civil society experts to advance the Network's work on the road to the AI Action Summit hosted by France in February.

San Francisco, California – Today the U.S. Department of Commerce and U.S. Department of State are co-hosting the inaugural convening of the International Network of AI Safety Institutes, a new global effort to advance the science of AI safety and enable cooperation on research, best practices, and evaluation. To harness the enormous benefits of AI, it is essential to foster a robust international ecosystem to help identify and mitigate the risks posed by this breakthrough technology. Through this Network, the United States hopes to address some of the most pressing challenges in AI safety and avoid a patchwork of global governance that could hamper innovation. The United States will serve as the inaugural chair of the International Network of AI Safety Institutes, whose initial members include Australia, Canada, the European Union, France, Japan, Kenya, the Republic of

Korea, Singapore, the United Kingdom, and the United States. Over the next two days in San Francisco, California, technical experts from member governments will be joined by leading AI developers, academics, civil society leaders, and scientists from non-Network governments to discuss key areas of collaboration on AI safety and lend their technical and scientific expertise to the Network's mission. The convening is structured as a technical working meeting that will address three high-priority topics that stand to urgently benefit from international coordination, specifically: (1) managing risks from synthetic content, (2) testing foundation models, and (3) conducting risk assessments for advanced AI systems. By bringing together the leading minds across governments, industry, academia, and civil society, we hope to kickstart meaningful international collaboration on AI safety and innovation, particularly as we work toward the upcoming AI Action Summit in France in February and beyond.

1) Adopting an aligned mission statement for the International Network of AI Safety Institutes. Launching the International Network of AI Safety Institutes is an essential step forward for global coordination on safe AI innovation. The Network will enable its members to leverage their respective technical capacity to harmonize approaches and minimize duplication of resources, while providing a platform to bring together global technical expertise. Ahead of the inaugural convening of the Network, all 10 initial members have agreed to a joint mission statement that reads in part: "The International Network of AI Safety Institutes is intended to be a forum that brings together technical expertise from around the world. Recognizing the importance of cultural and linguistic diversity, we aim to facilitate a common technical understanding of AI safety risks and mitigations based upon the work of our institutes and of the broader scientific community that will support international development and the adoption of interoperable principles and best practices. We also intend to encourage a general understanding of and approach to AI safety globally, that will enable the benefits of AI innovation to be shared amongst countries at all stages of development."

The International Network Members have also aligned on four priority areas of collaboration: pursuing AI safety research, developing best practices for model testing and evaluation, facilitating common approaches such as interpreting tests of advanced AI systems, and advancing global inclusion and information sharing. The United States AI Safety Institute (US AISI) will serve as the inaugural Chair of the International Network of AI Safety Institutes and Network members will discuss additional details of governance, structure, and meeting cadence at the convening. The Network will also discuss priorities and a roadmap for continued work toward the forthcoming AI Action Summit in Paris in February 2025 and beyond.

2) Announcing more than \$11 million in global research funding commitments to address the International Network's new joint research agenda on mitigating risks from synthetic content. With the rise of generative AI and the rapid development and adoption of highly capable AI models, it is now easier, faster, and less expensive than ever to create synthetic content at scale. Though there are a range of positive and innocuous uses for synthetic content, there are also risks that need to be identified, researched, and mitigated to prevent real-world harm – such as the generation and distribution of child sexual abuse material and non-consensual sexual imagery, or the facilitation of fraud and impersonation. To advance the state of the science and inform novel ways to mitigate synthetic content risks, the International Network of AI Safety Institutes has outlined a joint research agenda calling for urgent and actionable inquiry by the scientific community into key gaps in the current literature. Priority research topics include understanding the security and robustness of current digital content transparency techniques, exploring novel and emergent digital content transparency methods, and developing model safeguards to prevent the generation and distribution of harmful synthetic content. The International Network research agenda encourages a multidisciplinary approach, including technical mitigations as well as social scientific and humanistic assessments to identify problems and solutions.

In response to this agenda, government agencies and several leading philanthropies have committed a total of more than \$11 million (USD) to spur this vital research. The United States, through USAID, is designating \$3.8 million this fiscal year to strengthen capacity building, research, and deployment of safe and responsible AI in USAID partner countries overseas, including supporting research on synthetic content risk mitigation. Australia, through its national science agency, the Commonwealth Scientific and Industrial Research Organisation (CSIRO), is investing \$2.2 million (AUD) (\$1.42 million USD) annually in research aimed at identifying and mitigating the risks of synthetic content. The Republic of Korea will commit \$1.8 million (USD) annually for 4 years, totaling \$7.2 million (USD), toward research and development efforts on detecting, preventing, and mitigating the risks of synthetic content through the ROK AISI program. The John S. and James L. Knight Foundation is committing \$3 million (USD) to support scholarship aligned with the research agenda on mitigating risks from synthetic content. The AI Safety Fund (AISF), an independently administered non-profit collaboration between leading frontier AI developers and philanthropic partners, will contribute \$1 million (USD) in support of the research agenda, focusing on safeguards in the design, use, and evaluation of AI models and systems to reduce risks to security and public safety from model outputs. An additional \$250,000 (USD) in 2025 funding by the Omidyar Network has been

pledged in support of the research agenda. At the convening, the International Network will also discuss shared Network principles for mitigating synthetic content risks, including best practices and innovations for improving the security, reliability, privacy, transparency, and accessibility of generative AI. These inputs will help advance an aligned perspective from the members, which we hope to share at the forthcoming France AI Action Summit.

To support this work, US AISI is releasing the final version of its first synthetic content guidance report, NIST AI 100-4: Reducing Risks Posed by Synthetic Content , which identifies a series of voluntary approaches to address risks from AI-generated content like child sexual abuse material, impersonation, and fraud. This final version reflects input and feedback from an array of external experts and public comment solicited over the past several months.

(3) Methodological insights on multi-lingual, international AI testing efforts from the International Network of AI Safety Institutes' first-ever joint testing exercise. The International Network of AI Safety Institutes completed its first-ever joint testing exercise, led by technical experts from US AISI, UK AISI, and Singapore AISI. The Network conducted this exercise to explore methodological challenges, opportunities, and next steps for joint work that the Network can pursue to advance more robust and reproducible AI safety testing across languages, cultures, and contexts. This exercise raised key considerations for international testing, such as the impact that small methodological differences and model optimization techniques can have on evaluation results, and highlighted strategies for potentially mitigating these challenges. This exercise was conducted on Meta's Llama 3.1 405B to test across three topics – general academic knowledge, 'closed-domain' hallucinations, and multi-lingual capabilities – and will act as a pilot for a broader joint testing exercises leading into the AI Action Summit in Paris this February. The learnings from the pilot testing process will also lay the groundwork for future testing across international borders and evaluation best practices. (4) A joint statement on risk assessments of advanced AI systems, including a plan for advancing International Network alignment. Assessing the risks of advanced AI systems presents novel challenges, and it is central to the mission of the International Network of AI Safety Institutes to address these challenges and align on a framework for understanding the risks and capabilities posed by this technology. While recognizing that the science of AI risk assessment continues to evolve and that each Network member operates within its own unique context, the International Network of AI Safety Institutes agreed to establish a shared scientific basis for risk assessments , building on six key aspects outlined by the Network – namely, that risk assessments should be (1) actionable, (2) transparent, (3) comprehensive, (4) multistakeholder, (5) iterative, and (6) reproducible. This shared approach builds on commitments made in in the Bletchley Declaration and the Seoul Statement of Intent, as well as the progress made through the OECD, the G7 Hiroshima Process, the Frontier AI Safety Commitments, and other relevant international AI safety initiatives. At the convening, the International Network will solicit feedback and insight from members and gathered experts on how to operationalize a shared approach to risk assessments and build a roadmap for advancing global alignment and interoperability. (5) Establishing a new U.S. Government taskforce led by the U.S. AI Safety Institute to collaborate on research and testing of AI models to manage national security capabilities and risks. The Testing Risks of AI for National Security (TRAINS) Taskforce brings together experts from the Departments of Commerce, Defense, Energy, Homeland Security, as well as the National Security Agency (NSA) and National Institutes of Health (NIH) to address national security concerns and strengthen American leadership in AI innovation. The Taskforce will enable coordinated research and testing of advanced AI models across critical national security and public safety domains, such as radiological and nuclear security, chemical and biological security, cybersecurity, critical infrastructure, conventional military capabilities, and more. These efforts will advance the U.S. government imperative to maintain American leadership in AI development and prevent adversaries from misusing American innovation to undermine national security. More information can be found here in the press release . For press inquires, please reach out to usaisi [at] nist.gov

Super intelligence

Released November 20, 2024, Updated February 4, 2025

Was this page helpful?