OCTOBER 2024

The AI Safety Institute International Network

Next Steps and Recommendations

By Gregory C. Allen and Georgia Adamson

Overview

On November 21 and 22, 2024, technical artificial intelligence (AI) experts from nine countries and the European Union will meet for the first time in San Francisco. The agenda: starting the next phase of international cooperation on AI safety science through a network of AI safety institutes (AISIs). The United States, United Kingdom, European Union, Japan, Singapore, South Korea, Canada, France, Kenya, and Australia make up the initial members of the network, which was first **launched** by U.S. secretary of commerce Gina Raimondo at the May 2024 AI Seoul Summit. At the time of the launch, Italy and Germany were also potential members of the network, as signatories to the **Seoul Statement of Intent toward International Cooperation on AI Safety Science**, or Seoul Statement, the network's founding document. However, a September **announcement** by Raimondo and U.S. secretary of state Antony Blinken confirmed that Kenya would instead be the final member of the AISI International Network at this stage.

> *On November 21 and 22, 2024, technical artificial intelligence (AI) experts from nine countries and the European Union will meet for the first time in San Francisco. The agenda: starting the next phase of international cooperation on AI safety science through a network of AI safety institutes (AISIs).*

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

According to the Seoul Statement, the international network will serve to "accelerate the advancement of the science of AI safety" at a global level by promoting "complementarity and interoperability" between institutes and fostering a "common international understanding" of AI safety approaches. While the statement does not define specific goals or mechanisms for AISI collaboration, it suggests that they "may include" coordinating research, sharing resources and relevant information, developing best practices, and exchanging or codeveloping AI model evaluations. Now, in the months following the AI Seoul Summit, AISI network members must begin to articulate the objectives, deliverables, timelines, and avenues for cooperation that will put the promise of AISI cooperation into action.

> *In the months following the AI Seoul Summit, AISI network members must begin to articulate the objectives, deliverables, timelines, and avenues for cooperation that will put the promise of AISI cooperation into action.*

This paper examines next steps for developing the International Network of AI Safety Institutes from the Seoul Statement. It provides recommendations to members ahead of the inaugural network meeting in San Francisco this November and the **AI Action Summit** in Paris in February 2025. These recommendations fall in line with three key questions:

1. **Goals of collaboration: What is the AISI network trying to achieve and when?**
While there are many potential benefits to international collaboration, there are also real costs that should not be ignored. At a minimum, collaboration demands staff time, capacity, and possibly money from partners. The AISI network should therefore have clear goals for which type of international cooperation between safety institutes offers the maximum return on investment. These goals should be supported by specific priorities, deliverables, and timelines that steer the network's efforts toward a meaningful return on investment.
2. **Mechanisms of collaboration: What will the AISI network do and how will it work?**
The success of the network depends on how effectively its members can act upon shared goals. There are many different ways for the members to "collaborate," and not all of them are equally attractive. Network members should consider what the mechanisms of collaboration will be—for example, leadership structures, research exchanges, shared platforms, and annual conferences.
3. **International strategy: How will the AISI network fit into and engage with other international AI efforts?**
The AI governance landscape is increasingly crowded with international initiatives, including from the **Group of Seven** (G7), the **United Nations**, the **Organisation for Economic Co-operation and Development** (OECD), the **Global Partnership on AI** (GPAI), the **International Organization for Standardization** (ISO), and more. All of these demand time from a small (though growing) community of government staff from member countries who can credibly claim to have some expertise on AI governance and safety issues. AISI network members should be able to articulate how their grouping is different from these preexisting initiatives, how it will effectively engage with them (or not), and for what purpose.

This paper begins with background on the AISI network and explains its importance. Next, it offers an overview of network members' organizations and stated functions. It concludes with recommendations regarding nine further questions for developing the goals, collaboration mechanisms, and international strategy of the network.

Background

WHAT IS AI SAFETY AND WHY DOES IT MATTER?

As defined by the **Bletchley Declaration,** issued by attendees of the UK AI Safety Summit in November 2023, **AI safety is a scientific field** of research focused on evaluating, preventing, and mitigating risks from advanced AI systems. In this case, it refers narrowly to AI systems at or beyond the current state of the art. These risks can range from deepfakes to the use of AI for bioterrorism; new risks will emerge as AI's capabilities continue to evolve. Somewhat confusingly, other individuals and organizations may define AI safety more broadly to include lower-performing systems that are not operating at the technical frontier. Still others may or may not include issues around ethics and bias when using the term "AI safety." This paper's use of the term "AI safety" follows the **U.S. AI Safety Institute's example** of focusing exclusively on safety issues related to advanced AI systems.

AI safety science can be split into two main streams of research: technical safety, or improving the internal "machinery" of AI models; and process-based safety, or improving how people build, develop, and interact with AI models.

Technical AI safety focuses on understanding how the engineering and science behind AI models works, and how to make models perform reliably and in the scope of their intended use cases. These **three areas of research** are known as:

- **Assurance:** Understanding how a model makes decisions and why it behaves the way it does
- **Robustness**: Ensuring a model operates reliably under adverse contexts
- **Specification:** Designing a model that produces desired results as intended.

Meanwhile, **process-based safety** is concerned with the policies, practices, and procedures that surround AI. This stream of AI safety is more operational in nature. It focuses on how frontier AI developers, deployers, and users build, manage, and monitor AI models, including by evaluating models for capabilities, limitations, and risks, and documenting and reporting model information. It may also include processes that are implemented by the users of AI.

Beyond preventing adverse risks, AI safety serves to accelerate adoption and innovation by building public trust. As Elizabeth Kelly, director of the U.S. AI Safety Institute, said in a **CSIS interview**, "safety promotes trust, which promotes adoption, which drives innovation." AI safety boosts public trust by allowing people to pause, stop, or change course as needed.

A helpful analogy, one frequently used in the risk management sector, compares AI safety capabilities with the brakes on a car. At first consideration, the purpose of brakes seems obvious and narrow: to make the car go slower. However, the existence of brakes also allows cars to go faster. As a thought experiment, imagine how fast drivers would be willing to go if no car came equipped with any brakes. How easy would it be to avoid a crash or turn a corner if drivers could never change their speed? How might one pause to change tires or fix a problem? Navigating such scenarios would almost certainly be a disaster. Even with speed limit regulations in place, a world without brakes would be a world in which drivers went much, much slower.

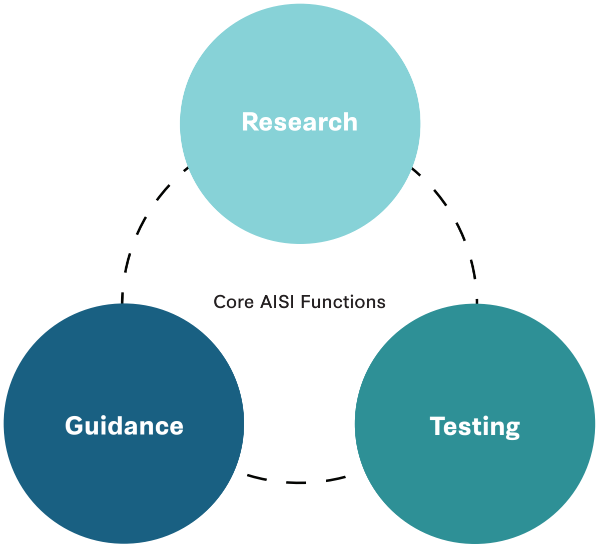
Policymakers should approach AI safety with this parallel in mind. Like the brakes of a car, building technical and management capabilities for AI can help boost confidence in the technology and ultimately accelerate the pace of adoption and innovation.

WHAT ARE AI SAFETY INSTITUTES AND WHAT WILL THEY DO?

Since 2023, governments around the world have mobilized around AI's rapidly growing capabilities and potential risks. As part of this effort, several governments have launched AI safety institutes, publicly funded research institutions focused on mitigating risks from the frontier of AI development. AISIs provide governments with in-house technical expertise and organizational capacity to evaluate and monitor cutting-edge AI models for risks to public and national security.

AISIs have been tasked by governments with a wide-ranging mandate to address the complex challenges posed by advanced AI systems. They will perform foundational technical research, develop guidance for the public and private sectors, and work closely with companies to test models before deployment. While it is unusual for a single government entity to tackle all three of these functions at once, the breakneck speed of AI development and the staggering number of open questions in the field of AI safety research mean that governments require in-house capacity on each of them. **According to Kelly**, it is important that these three functions–research, testing, and guidance–reinforce each other to form a "virtuous" cycle (Figure 1):

Figure 1: AISI Core Functions



Source: "The U.S. Vision for AI Safety: A Conversation with Elizabeth Kelly, Director of the U.S. AI Safety Institute," CSIS, July 31, 2024, https://www.csis.org/analysis/us-vision-ai-safety-conversation-elizabeth-kelly-director-us-ai-safety-institute; and "The United States Artificial Intelligence Safety Institute: Vision, Mission, and Strategic Goals," U.S. Artificial Intelligence Safety Institute, May 21, 2024, https://www.nist.gov/system/files/documents/2024/05/21/AISI-vision-21May2024.pdf.

To keep pace with the cutting edge of AI safety research, AISIs have prioritized the hiring of technical staff and opened offices in cities with deep pools of AI talent like San Francisco. In addition to developing expertise internally, AISIs aim to cultivate a robust ecosystem of AI safety researchers in labs, industry, and **academia** through their guidance on best-in-class evaluation methods.

AISIs are engaging a wide range of stakeholders on each of their core functions. Far from fearing the launch of AISIs worldwide, firms and universities engaged in advanced AI have called for governments to increase their capacity to perform AI research, conduct testing, and issue guidance. Earlier **this year**, top U.S. AI companies such as Google, Microsoft, Anthropic, and Amazon joined the **U.S. AISI Consortium** (AISIC) as part of its inaugural cohort of members. AISIC is composed of over 200 organizations from across the private sector, academia, civil society, and government and facilitates collaboration on AI safety research and evaluations. Members are expected to contribute to one of **nine key areas of guidance**, reproduced verbatim below:

1. Develop new guidelines, tools, methods, protocols, and best practices to facilitate the evolution of industry standards for developing or deploying AI in safe, secure, and trustworthy ways
2. Develop guidance and benchmarks for identifying and evaluating AI capabilities, with a focus on capabilities that could potentially cause harm
3. Develop approaches to incorporate secure-development practices for generative AI, including special considerations for dual-use foundation models, including:
    - Guidance related to assessing and managing the safety, security, and trustworthiness of models and related to privacy-preserving machine learning
    - Guidance to ensure the availability of testing environments
4. Develop and ensure the availability of testing environments
5. Develop guidance, methods, skills, and practices for successful red-teaming and privacy-preserving machine learning
6. Develop guidance and tools for authenticating digital content
7. Develop guidance and criteria for AI workforce skills, including risk identification and management; test, evaluation, validation, and verification (TEVV); and domain-specific expertise
8. Explore the complexities at the intersection of society and technology, including the science of how humans make sense of and engage with AI in different contexts
9. Develop guidance for understanding and managing the interdependencies between and among AI actors along the lifecycle.

Note that while these nine areas of guidance overlap with the nine core functions of an AI safety institute identified in Section 4 of this paper, they do not cover the full breadth of AISIs' operations. As Section 4 will discuss, AISIs perform functions such as forming consortia of AI researchers, stakeholders, and experts and promoting the international adoption of AI safety guidelines that are outside the scope of the AISIC.

In August, OpenAI chief executive officer Sam Altman **stated** that his company has been working closely with the U.S. AISI on an agreement to provide early access to its next foundation model for safety testing and evaluations. OpenAI is not alone in providing the U.S. AISI access to its models for testing. Director Kelly said that the institute has “commitments from all of the leading frontier model developers to work with them on these tests.” These commitments demonstrate that leading companies understand the need for AI safety research and recognize the important role that the U.S. AISI has to play. While critics have questioned how industry will balance competition and safety, AISIs are free from the financial self-interest which has **caused some to question** the adequacy of private AI safety efforts in the past.

On October 21, top AI developers including Amazon, Meta, Microsoft, and OpenAI signed **a letter** to Congress calling on lawmakers to authorize the U.S. AISI before the end of the year. The letter, which was led by Americans for Responsible Innovation and the Information Technology Industry Council (ITI), states that "[a]s other nations around the world are establishing their own AI Safety Institutes, furthering NIST's ongoing efforts is essential to advancing U.S. AI innovation, leadership, and national security." "Authorizing legislation, and the accompanying necessary resources,” it argues, “will give much needed certainty to NIST's role in AI safety and reliability.”

The letter echoes similar **calls** for Congress to authorize the AISI by Scale AI Founder and CEO Alexandr Wang earlier in October, as well as a **letter** from top AI companies to establish the AISI on a statutory basis in July. The July letter, also published by Americans for Responsible Innovation and ITI, argues that authorizing the AISI "provides a venue to convene the leading experts across industry and government to contribute to the development of voluntary standards that ultimately assist in de-risking adoption of AI technologies.” It’s not just the biggest companies that stand to benefit from the U.S. AISI—crucially, the letter argued that the institute may level the playing field for enterprises that use or develop AI but are unable to perform robust testing and evaluation in-house due to their size or the technical ability of their staff.

While the concept of a government organization that works closely with AI companies on safety is still new, history shows that this kind of arrangement between government and industry can be highly successful. One good example is the **National Highway Traffic Safety Administration** (NHTSA), a U.S. federal agency that performs safety tests of new motor vehicle models for manufacturers. Established **in the 1970s** to reduce accidents and deaths by encouraging manufacturers to produce safer vehicles, NHTSA led what has become today an industry **standard** of crash testing and rating vehicles out of five stars according to their safety. Some 50 years since its launch, NHTSA continues to perform crash tests and produce star ratings, as well as issue government safety ratings, safety information, and best practices.

NHTSA is a useful model of a third-party government arbiter that has produced substantial win-win results for the public and for companies. The administration’s rating system lowers costs to consumers by supplying accurate, reliable, and simple safety information for free. Meanwhile, companies are incentivized to adopt new and better safety measures into their vehicles. As NHTSA’s acting administrator **has stated**, “[o]ur 5-Star Safety Ratings system continues to give Americans the information they need to choose the vehicle that’s right for them. The program also encourages vehicle manufacturers to incorporate advanced vehicle safety technologies into more makes and models, ultimately reducing injuries and deaths on America's roads." Because **safety is a selling point for customers**, **most** of the United States' manufacturers willingly sign up for the NHTSA's 5-star system and use the results in **advertising** new vehicle models.

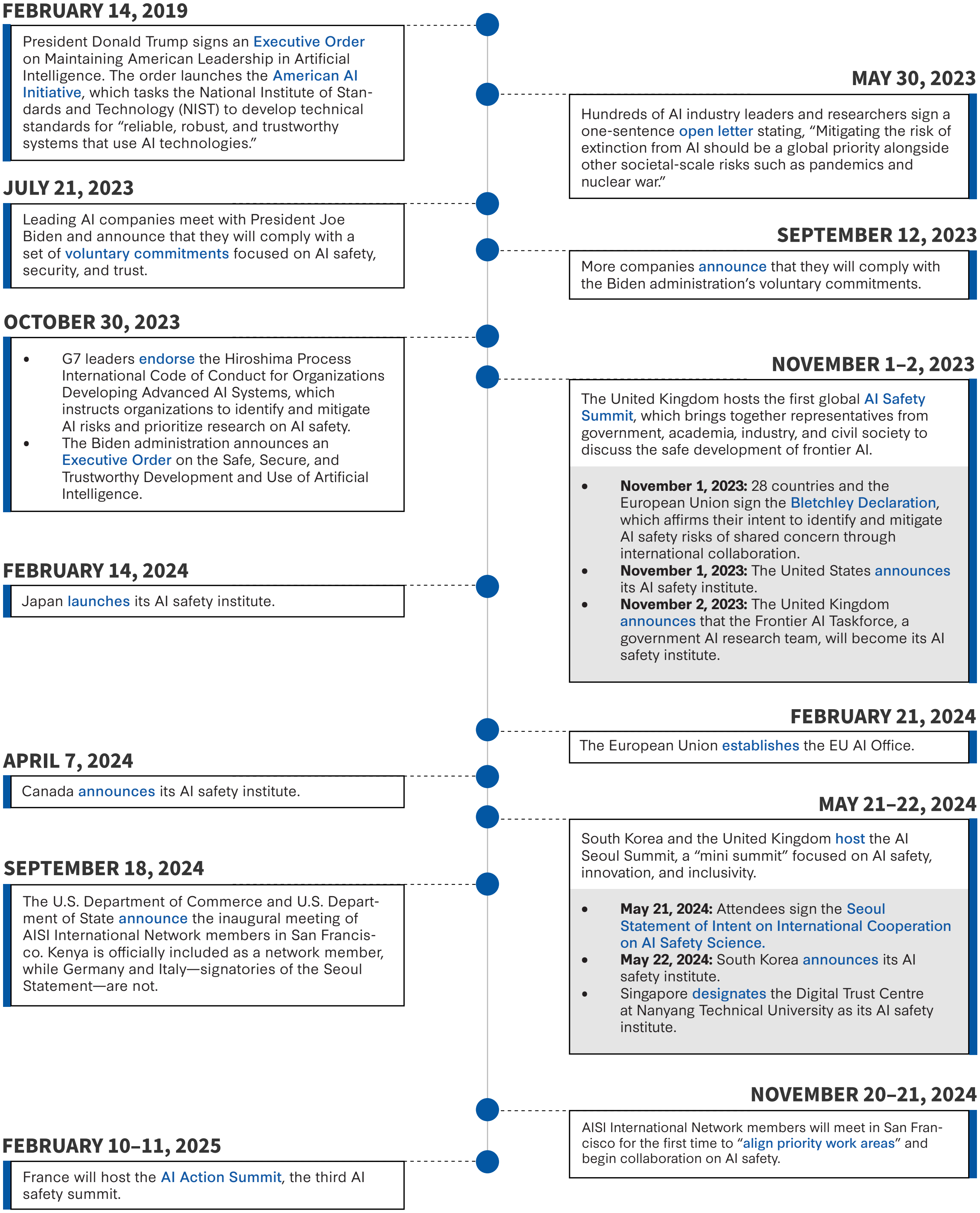
As AISIs mature organizationally, they could fulfill a similar arbiter role for AI models as the NHTSA has for motor vehicles. As has been the case with motor vehicles, testing AI models could lead to innovation in which safety is a key competitive feature. AI companies could communicate to customers that their model has passed AISI testing and evaluations, which could in turn help to build public trust and make AI models with higher safety standards more commercially competitive among consumers. Top frontier AI developers' willingness to work with the U.S. AISI on testing their models before deployment is a good first step to making safety a key feature of AI industry standards, as the NHTSA has done with the U.S. motor vehicle industry over the last 50 years.

TIMELINE OF AI SAFETY INSTITUTES

The first AISIs were announced last year, with the **United States** and **United Kingdom** launching initiatives at the UK AI Safety Summit in November 2023. **Japan**, **Singapore**, and the European Union's **EU AI Office** followed in early 2024. Since then, **Canada** and **South Korea** have revealed plans for their own AISIs. The inclusion of France, Kenya, and Australia in the AISI network suggests that more institutes are still to come. For instance, in May French research institutions Laboratoire National de Métrologie et d'Essais (LNE) and National Institute for Research in Digital Science and Technology (Inria) **announced a partnership** to set up an "AI Evaluation" program that will advance research and the development of testing and evaluation methods for general-purpose AI models at the national level. While this program has not yet been named as an official AI safety institute for France, an announcement may take place at the AI Action Summit in France in February 2025, similar to the announcement made by South Korea at the AI Seoul Summit in May.

The AISI International Network marks a logical next step in a series of recent bilateral agreements between institutes. In April 2024, the United States **signed** a memorandum of understanding with the United Kingdom for close collaboration between institutes and established a **dialogue** with the EU AI Office to jointly develop evaluation tools for AI models. Meanwhile, the United Kingdom, for its part, has established additional partnerships with **Canada** and **France** on AI safety, and the European Union and Japan have indicated **future cooperation** between safety institutes in the coming months.

Figure 2: Timeline of Major Events in AI Safety Since 2019



FEBRUARY 14, 2019

President Donald Trump signs an Executive Order on Maintaining American Leadership in Artificial Intelligence. The order launches the American AI Initiative, which tasks the National Institute of Standards and Technology (NIST) to develop technical standards for “reliable, robust, and trustworthy systems that use AI technologies.”

MAY 30, 2023

Hundreds of AI industry leaders and researchers sign a one-sentence open letter stating, “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”

JULY 21, 2023

Leading AI companies meet with President Joe Biden and announce that they will comply with a set of voluntary commitments focused on AI safety, security, and trust.

SEPTEMBER 12, 2023

More companies announce that they will comply with the Biden administration’s voluntary commitments.

OCTOBER 30, 2023

- G7 leaders endorse the Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, which instructs organizations to identify and mitigate AI risks and prioritize research on AI safety.
- The Biden administration announces an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

NOVEMBER 1–2, 2023

The United Kingdom hosts the first global AI Safety Summit, which brings together representatives from government, academia, industry, and civil society to discuss the safe development of frontier AI.

- **November 1, 2023:** 28 countries and the European Union sign the Bletchley Declaration, which affirms their intent to identify and mitigate AI safety risks of shared concern through international collaboration.
- **November 1, 2023:** The United States announces its AI safety institute.
- **November 2, 2023:** The United Kingdom announces that the Frontier AI Taskforce, a government AI research team, will become its AI safety institute.

FEBRUARY 14, 2024

Japan launches its AI safety institute.

FEBRUARY 21, 2024

The European Union establishes the EU AI Office.

APRIL 7, 2024

Canada announces its AI safety institute.

MAY 21–22, 2024

South Korea and the United Kingdom host the AI Seoul Summit, a “mini summit” focused on AI safety, innovation, and inclusivity.

- **May 21, 2024:** Attendees sign the Seoul Statement of Intent on International Cooperation on AI Safety Science.
- **May 22, 2024:** South Korea announces its AI safety institute.
- Singapore designates the Digital Trust Centre at Nanyang Technical University as its AI safety institute.

SEPTEMBER 18, 2024

The U.S. Department of Commerce and U.S. Department of State announce the inaugural meeting of AISI International Network members in San Francisco. Kenya is officially included as a network member, while Germany and Italy—signatories of the Seoul Statement—are not.

NOVEMBER 20–21, 2024

AISI International Network members will meet in San Francisco for the first time to “align priority work areas” and begin collaboration on AI safety.

FEBRUARY 10–11, 2025

France will host the AI Action Summit, the third AI safety summit.

Why The AISI International Network Matters

The AISI International Network is important for several reasons:

- **The network provides a much-needed venue for building international consensus on definitions, procedures, and best practices around AI safety.** The science of evaluating AI models is a nascent yet vital field of research that underpins global efforts to develop safe and responsible AI. Currently, these efforts are limited by a lack of consensus on key definitions (for instance, what constitutes a "frontier" AI model or a "secure" system) and on the steps involved in testing, evaluation, and monitoring procedures.

  International consensus would increase regulatory interoperability, or the degree to which different domestic regulatory systems can smoothly interface and interact. Interoperability allows for the even implementation of international AI governance efforts. One such effort is the **G7 Hiroshima AI Process Code of Conduct,** which calls for "robust" and "trustworthy" AI systems but lacks technical definitions of the terms. Shared definitions would help create a common measuring stick by which regulators gauge these characteristics. Countries could choose policy options along such a ruler based on their risk tolerance for given AI applications. In this example, governments would require different levels of robustness and trustworthiness along the same underlying scale, as is the case for safety in the automobile and aviation industries. A common understanding of AI safety concepts would help clarify the steps countries must take to honor the G7 code of conduct and other international commitments.

  In this way, interoperability based on common definitions, procedures, and best practices can help to facilitate trade in the future. As a previous **CSIS paper** argued, fragmented legal frameworks that require company compliance with many different obligations can create technical barriers to the free flow of goods and services. Diverging regulatory approaches that require companies to demonstrate that a product is "safe" according to 10 different metrics from 10 different jurisdictions, for instance, is not only highly inefficient but often prohibitively costly. Instead, the AISI International Network could serve as one venue in which to develop a coherent language around AI safety, helping to lower future potential barriers to trade.

- **International collaboration will help governments achieve economies of scale in AI safety research.** Thus far, AISIs have cooperated on a bilateral basis, which, while useful, can limit the impact and scope of AI safety efforts. By sharing priorities, resources, and expertise through a multilateral configuration, the AISI International Network aims to be more than the sum of its parts. AISIs can contribute strategically to the goals of the network by coordinating roles and responsibilities, de-duplicating research and therefore saving time, capacity, and money.

- **The network offers an opportunity to extend U.S. leadership in global AI governance.** The United States has already demonstrated significant leadership in AI safety by being one of the first to launch its AISI in 2023 and by spearheading the AISI network initiative in 2024. It should maintain this leadership going forward with the view that the network will help shape global AI safety practices that will predominantly affect U.S. companies.

This is important for not only setting safety norms at home, but also advocating for U.S. interests abroad. Consider, for instance, the EU AI Act: while the first wave of the act **came into force** on August 1, the requirements for developers of frontier AI models above **10^25** floating operation points (FLOPS) of compute power have yet to be defined. Rather, the EU AI Office—the European Union's representation to the AISI International Network—is tasked with **developing codes of practice** for the developers of these models, almost all of which are **U.S. companies**.

According to **Article 56** of the AI Act, the EU AI Office must develop codes of practice for frontier AI companies to identify, assess, manage, and report "systemic" risks by May 2, 2025. To meet this tight deadline, it may look to the work of the AISI International Network if it deems it sufficiently mature to draw upon. Having a seat at the same table as the EU AI Office is therefore a valuable opportunity to help develop safety norms that the European Union may apply to U.S. companies. Even if the European Union ultimately decides to develop its codes of practice alone, the network will still provide the United States with a direct line of communication to the EU AI Office for articulating AI safety best practices in the future.

Overview of AISI Network Members

It is still early days for AI safety institutes, both as organizations and as concepts. Members of the AISI International Network are highly varied in their organizational maturity, which can be expected given that most are only months old. Even the U.S. AISI, one of the most established institutes, was announced only in November 2023 and became **operational** in early 2024. Other AISIs, such as those of Japan, Singapore, South Korea, and the European Union, are still in the process of hiring and setting out the priorities of their institutes, according to public documents and conversations by CSIS with officials. Still other network members, like Kenya and Australia, have yet to clearly state whether their governments will even establish an AISI.

Nevertheless, established AISIs report strong similarities in funding and staff size thus far. As Table 1 illustrates, the annual budgets of network members currently hover around $10 million, with some notable exceptions. First, the UK AISI is already an outlier with a budget of approximately £50 million ($65 million) per year, according to CSIS sources. Second, the United States' fiscal year 2025 budget requests an increase of **$47.7 million** for investment into the U.S. AISI and the advancement of AI research, standards, and testing in line with President Biden's October 2023 **AI executive order**, which, if approved, would greatly boost the average network budget. Finally, an **announcement** by the Canadian government in April pledges C$50 million (approximately US$36 million) for a Canadian AISI, though the funding period is unspecified.

Public statements and private conversations between CSIS and government officials reveal that staff sizes will also be comparable between institutes. More established AISIs currently employ approximately 20 to 30 staff, most of whom are technical experts. Private conversations with CSIS indicate that the EU AI Office's **AI safety unit**, which will fulfill most of the same functions as an AISI (Table 2), will likely hold approximately 50 staff members.

Table 1: Organizational Overview of AISI Network Members

	United States	United Kingdom	European Union	Japan	Singapore	South Korea	Canada	France	Kenya	Australia
Established	February 2024	November 2023	May 2024	February 2024	May 2024	May 2024 (Announced)	April 2024 (Announced)			
Name of Organization	**US AISI**	**UK AISI**	**EU AI Office**	**Japan AISI**	**Singapore AISI**	**Korea AISI**	**Canada AISI**			
Housed Under	National Institute of Standards & Technology	Department for Science, Innovation & Technology	Directorate General for Communications Networks, Content and Technology.	Information-Technology Promotion Agency	Digital Trust Centre	Electronics and Telecommunications Research Institute				
Funding (USD & Foreign Currency)	**$10 million** (FY24)	> **$65 million/yr** (>£50 million/yr 2024-2030)	**$51 million** (€46.5 million) (Funding period unknown)		**$7.5 million/yr** (S$10 million/year) (2023-2027)	**$7.2-14.4 million/yr** (W10-20 billion/yr) (Tentative, starting 2025)	**$36.5 million** (C$50 million) (Funding period unknown)			
Staff	c.20 (current core staff)	c.20 (current core staff)	c.50 (planned, AI safety unit)	c. 23 (current staff)		Minimum 30 staff (planned, budget pending)				
Public List of Functions	US AISI Vision, Mission, and Strategic Goals	Introducing the AI Safety Institute	Tasks of the AI Office	AISI's Tasks	Initial Research Areas					
Published Research or Guidelines	Managing Misuse Risk for Dual-Use Foundation Models	See website		See website	Model AI Governance Framework for Generative AI					

Legend	No public statement	Public Information

Source: Public statements from AISI network members and relevant government officials and bodies.

AISI network members also intend to fulfill similar functions. Based on a document review of all public statements from AISIs and relevant government officials, this paper provides a list of the nine areas of AI safety in which institutes may operate (see Table 2). These functions are:

1. Performing (technical) research on AI safety tools
2. Developing and disseminating evaluation tools and products
3. Testing and evaluating AI systems
4. Publishing AI safety standards and guidelines
5. Disseminating AISI research and guidelines to the public
6. Forming consortia of AI researchers, stakeholders, and experts
7. Promoting the international adoption of AI safety guidelines
8. Investigating infringements of domestic regulations
9. Encouraging domestic innovation in AI

Table 2 demonstrates that most AISI network members will principally focus on the first seven of these nine functions, with notably only the European Union performing a regulatory role as part of the EU AI Office. This overlap between network members' stated functions points to a strong basis for collaboration between AISIs.

It also shows that some institutes have already begun to produce work related to their stated functions. Some deliverables predate the AISI, such as the Japanese Ministry of Economy, Trade and Industry's AI Business Guidelines, but have been incorporated and built upon by current AISI efforts. Others are novel efforts by institutes since their launch, such as the U.S. AISI's guidance for **Managing Misuse Risk for Dual-Use Foundation Models**, and the UK AISI's **Inspect** and Singapore's **Project Moonshot**, two testing and evaluation toolkits for large language models (LLMs).

Table 2: Overview of AISI Network Members' Stated Functions

	United States	United Kingdom	European Union	Japan	Singapore	Canada	South Korea	France	Kenya	Australia
Perform research on AI safety tools (technical)	●	●	◔	◑	◔	○	○	○	○	○
Build and release evaluation tools and products	◑	●	◑	◔	●	○	○	○	○	○
Test and evaluate AI systems	●	●	◑	◑	◑	○	○	○	○	○
Publish AI safety standards and guidelines	●	◔	◑	●	●	○	○	○	○	○
Inform policymakers and the public on research and safety guidelines	●	●	◑	◑	●	○	○	○	○	○
Form consortia of AI researchers, stakeholders, & experts	●	◑	●	◔	◔	○	○	○	○	○
Promote international adoption of AI safety guidelines	◑	◔	◑	◑	◑	○	○	○	○	○
Investigate infringements of domestic regulations	○	○	◑	○	○	○	○	○	○	○
Encourage domestic innovation in AI	○	○	◑	○	◑	○	○	○	○	○

Legend	○ No public statement	◔ Function not publicly stated but collaboration likely	◑ Function publicly stated	● Function operationalized

Source: Public statements from AISI network members and relevant government officials and bodies.

It is worth noting, however, that while institutes share many similarities in funding, size, and functions, they are housed under different kinds of public bodies. Several institutes are located within government agencies focused on technological innovation and standards, including the U.S. National Institute of Standards and Technology (NIST); the UK Department for Science, Innovation and Technology (DSIT); and the Japanese Information Technology Promotion Agency (IPA). Others are housed in government-funded research organizations, like the South Korean Electronics and Telecommunications Research Institute (ETRI) and the Singaporean Digital Trust Centre, itself a part of Nanyang Technological University. Finally, as Table 2 illustrates, the EU AI Office has the largest set of functions as an institution that promotes innovation, research, and regulatory compliance to the EU AI Act. The different kinds of home institutions in which AISIs are housed may have implications for the focus and capacity of different network members, and therefore the strengths that each member may bring to the network.

Questions and Recommendations

Similarities between AISI network members in terms of funding, size, and stated functions are a strong foundation for international cooperation on AI safety. However, ensuring that the AISI International Network maintains momentum requires translating the high-level Seoul Statement into a concrete set of priorities, deliverables, and timelines. To do so, this paper poses the following nine questions and recommendations to network members:

GOALS OF COLLABORATION: WHAT IS THE AISI NETWORK TRYING TO ACHIEVE AND WHEN?

1. **What areas of collaboration should the AISI network prioritize in the near term?**

   Recommendation: The AISI International Network does not have the capacity or resources to effectively collaborate on every domain of AI safety. For some domains, such as sharing sensitive information about models, AISIs may even face legal limitations to collaboration. Rather than spreading finite resources thinly in an effort to achieve everything all at once, network members should first focus on executing a few specific projects well. These should be attainable in the near future to demonstrate continued momentum from the AI Seoul Summit.

   When selecting priority areas, members should consider areas with the greatest overlap in AISI's functions, capacity, and expertise, and deliverables that are both impactful and realistic. To start, they should establish a research agenda for the network's technical and guidance safety work going forward. This will help to set the scope of the network's efforts and to keep members on track as they and the network mature. As discussed in this paper's recommendation to Question 3, the AISI network conference in November may be a good place to set and present this agenda to the public.

   In the medium term, network members should look to develop a common, evidence-based approach to AISIs' testing and evaluation methodologies. While not all AISIs may necessarily have the same requirements for assessing models, they should at least have a common understanding of what methodologies such as "red teaming" comprise. Developing a consensus on testing and evaluation methods would help to deconflict and de-duplicate efforts between AISIs and to facilitate other areas of collaboration in the future, such as promoting safety guidelines or developing joint evaluation tools. If the AISI network can start by ensuring that AISIs all speak the same language in AI safety, more elaborate collaboration projects can take place.

2. **What deliverables should the AISI network aim to produce?**

   Recommendation: Although the AISI network is very new, members should still consider what the end products of their collaboration might be. One of the first deliverables that the network could produce is a clear statement of its intended goals, functions, research agenda, and mechanisms of collaboration that builds on the Seoul Statement. In as much detail as possible, the statement should articulate the mission of the network, its intended scope of work, and how it will relate to other international organizations working on AI. Network members may also consider developing a comprehensive list of the specific risks that they will test. This statement would not only help network members set the agenda for collaboration, but it would also help external governments and organizations to understand the value of the AISI network and how the network can support their efforts.

3. **What are some key dates for these deliverables?**

   Recommendation: There are two big international events related to AI safety on the horizon that offer some initial deadlines for AISI network deliverables. First, the November 2024 San Francisco convening is an obvious date to publicly initiate international collaboration on AI safety. In September, the U.S. Department of Commerce and U.S. Department of State **announced** that "the goal of this convening is to kickstart the Network's technical collaboration ahead of the AI Action Summit in Paris in February 2025," starting with aligning "on priority work areas for the Network," as the recommendation above supports. The February summit, therefore, is an important second date for network deliverables. The AI Action Summit will be the third of its kind since the UK AI Safety Summit last year and offers a high-profile, public venue in which to showcase the AISI network and its work. These two events–in November 2024 and February 2025–are mere moments away in the context of international collaboration. If AISI members can capitalize on their opportunities, however, they could significantly contribute to the network's mission of accelerating AI safety science.

MECHANISMS OF COLLABORATION: WHAT WILL THE AISI NETWORK DO AND HOW WILL IT WORK?

4. **How will network members collaborate?**

   Recommendation: AISIs should aim to have a regular cadence of meetings, perhaps every six months, to sustain momentum and keep collaboration moving forward. AISIs could collaborate through any number of venues, including research exchanges, annual conferences, shared digital platforms, and more. Network members will likely use a mix of these and other venues in different combinations as the network matures over time. To start, research exchanges between AISIs may be one of the first mechanisms of collaboration given that it is relatively inexpensive.

5. **Will network members specialize in their work, or will they share equal responsibilities?**

   Recommendation: It would be premature to assign specific responsibilities to AISI network members today given that most are only months old, if established at all. However, members should consider the benefits and drawbacks of different organizational structures as the network develops. Currently, AISI network members share equal responsibilities by default. While this can be useful for promoting equal participation and accountability from members, it can also add unnecessary costs to collaboration. If each member were to take charge on a different project, for instance, the network could risk losing time, capacity, and focus. This kind of structure could also place undue pressure on the capacity and expertise of each of the AISIs to contribute before they are ready.

Instead, the AISI network may consider leveraging each member's comparative advantages in expertise, capacity, and funding. Those that are most able to contribute to projects, for instance, should be able and incentivized to do so, as is discussed in Question 7. For now, more mature AISIs like those of the United States, the United Kingdom, and Singapore could have greater responsibilities within the network while other members, such as Kenya or Australia, contribute through more specialized ways. These roles could shift over time as AISIs mature, however.

6. **Will the AI safety summits continue to serve as the principal international venue for AISIs and the AISI network?**

   Recommendation: Since the first AISIs were announced at Bletchley Park in November 2023, AISIs have been closely tied to the AI safety summits. However, the summit series is steadily shifting its focus from AI safety to AI adoption and innovation; in May, the AI Seoul Summit placed AI innovation and inclusiveness firmly on the agenda. The next summit, the AI Action Summit in February 2025, will reportedly include AI safety as only one of five topic areas.

   Nevertheless, a shift in focus does not mean that summits are not a good international venue for AISIs and the AISI network. In fact, it may make it an even better venue for helping to shift the rhetoric around AI safety from "doom and gloom" to "safety for trust, adoption, and innovation"—a far more politically salient message. This paper therefore recommends that AISIs and the AISI network continue to use the AI safety summits as a high-profile international venue for their efforts for as long as the summit series continues.

7. **What will the network's leadership and voting structure look like?**

   Recommendation: Currently, the AISI network has a horizontal leadership and consensus or opt-in only voting structure by default. Given that the Seoul Statement makes no indication of leadership and voting structure, however, network members are open to consider different possibilities and their trade-offs. For example, a consensus-based structure can help to foster good intentions for international cooperation, but it can also make it challenging to take meaningful collective action. Similarly, having just one member serve as a leader may seem unfair, but a rotating leadership structure can be ineffectual and prioritize the interests of that country (or bloc) for that period.

   The network's leadership and voting structures need not be zero sum, however. In the long run, members' representation within the network should be proportionate to their contributions; those that invest more time, money, expertise, and resources should be rewarded with a greater say in its direction. This means that the U.S. and UK AISIs would likely be rewarded with leadership of the network due to their organizational capacity. The United States, for its part, should aspire to lead the AISI network, as discussed in the third section of this paper. Rather than merely insisting on leading, however, it should commit the resources and time that positions it to *deserve* to lead. Leadership should be earned based on the scale of meaningful contributions to the field of AI safety science, a structure that also incentivizes on other network members to participate and invest more into AI safety and the AISI network as well.

INTERNATIONAL STRATEGY: HOW WILL THE AISI NETWORK FIT INTO AND ENGAGE WITH OTHER INTERNATIONAL AI EFFORTS?

8. **How will the network be different from and engage with other international organizations working on AI issues, such as the ISO, G7, United Nations, GPAI, or OECD?**

   Recommendation: Just as one of the objectives of the AISI network is to avoid duplicating work between AISIs, the network itself should avoid duplicating the work of other international organizations. Considering how the AISI network will fit into the broader landscape of these organizations from the start will help members think more strategically about what role this forum plays on the global AI governance stage.

   To do this, the AISI network should emphasize its unique position to provide technical expertise and capacity to governments working on wider AI governance efforts. In the past year alone, numerous government initiatives have been launched to ensure responsible frontier AI development, including the Biden administration's **AI executive order**, the **EU AI Act**, the **G7 Hiroshima AI Process Code of Conduct**, and the March 2024 **UN resolution on AI**. These initiatives, though commendable, are often staffed by diplomats who lack the depth of in-house technical expertise that the AISI network has demonstrated an ability to amass. It is this expertise that could turn what are currently high-level principles and frameworks into practical implementation for developers.

   For instance, the G7's code of conduct instructs developers to "identify, evaluate, and mitigate risks across the AI lifecycle," but provides little guidance as to how. While the G7 has partnered with the **OECD** to develop this level of specificity for the code of conduct, it would greatly benefit from the testing and evaluation tools that the Seoul Statement names as one of the potential areas for collaboration within the AISI International Network. Network members should consider how to engage with other international organizations' AI efforts with these synergies in mind.

9. **Will the network remain a closed group of high-capacity countries, or will it be open to any country that wants to join?**

   Recommendation: The AISI International Network was born out of recognition that AI risks do not stop at national borders. It therefore makes sense that the network should be open to more members that want to join. A wider membership would help to build international agreement on AI safety science and potentially to continue to reach economies of scale on AI safety institutes. Furthermore, incorporating more developing countries' perspectives early on–either through full membership or agreements–could bring new insight into AI safety risks that current AISI network members and their companies may have missed.

   However, network members will have to consider the serious trade-offs between expanding the network's membership and diluting its current nimbleness and consensus as a small group. Even countries or blocs that are closely aligned in values to current members may lack the technical expertise to meaningfully contribute to the network, thus raising the costs of collaboration and possibly reducing its impact.

One way to address this could be requiring prospective members to demonstrate their ability to meaningfully contribute to the network—such as through a minimum degree of expertise and capacity—before they can join. The purpose here is not to make the AISI network into an elite club, but to recognize that the network's goal of accelerating AI safety science cannot be realistically achieved by expanding membership to everyone who wants it. The AISI network could consider partnership programs with other international organizations like GPAI, the OECD, or the Group of 20 (G20) in order to collaborate with interested countries that do not necessarily have the depth of AI safety expertise to join the network. Such partnerships could help to foster wider international cooperation on AI safety and engage more developing countries on the AISI network's efforts in particular.

Conclusion

The AISI International Network marks a significant next step in global AI safety efforts. The network provides an opportunity to build international consensus on definitions, procedures, and best practices around AI safety; reach economies of scale in AI safety research; and extend U.S. leadership in international AI governance. The similarities between currently established AISIs in terms of size, funding, and functions provide a strong basis for cooperation, though network members must be aware of the different institutions in which different AISIs are housed.

While the Seoul Statement is a good start for multilateralizing cooperation between AISIs, network members must now decide how to turn intent into action. At the November convening in San Francisco, they should strive to set the network's goals, mechanisms, and international strategy in preparation for the AI Action Summit in February 2025. In doing so, they must ask tough questions, including about priorities, leadership, and membership. ■

Gregory C. Allen *is the director of the Wadhwani AI Center at the Center for Strategic and International Studies (CSIS) in Washington, D.C.* ***Georgia Adamson*** *is a research associate with the Wadhwani AI Center at CSIS.*

This report is made possible through the generous support of Microsoft.

This report is produced by the Center for Strategic and International Studies (CSIS), a private, tax-exempt institution focusing on international public policy issues. Its research is nonpartisan and nonproprietary. CSIS does not take specific policy positions. Accordingly, all views, positions, and conclusions expressed in this publication should be understood to be solely those of the author(s).