THE FUTURE SOCIETY

NOVEMBER 2024

Weaving a Safety Net

Key Considerations for How the AI Safety Institute Network Can Advance Multilateral Collaboration

Frank Ryan, Niki Iliadis, George Gor

THE
FUTURE
SOCIETY

Weaving a Safety Net:
Key Considerations for How the AI Safety Institute Network Can Advance Multilateral Collaboration

Cite as:

Frank Ryan, Niki Iliadis, and George Gor, "Weaving a Safety Net: Key Considerations for How the AI Safety Institute Network Can Advance Multilateral Collaboration" (The Future Society, 2024).

Contact: info@thefuturesociety.org

© 2024 by The Future Society. Design by Vilim Pavlović.
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

THE FUTURE SOCIETY

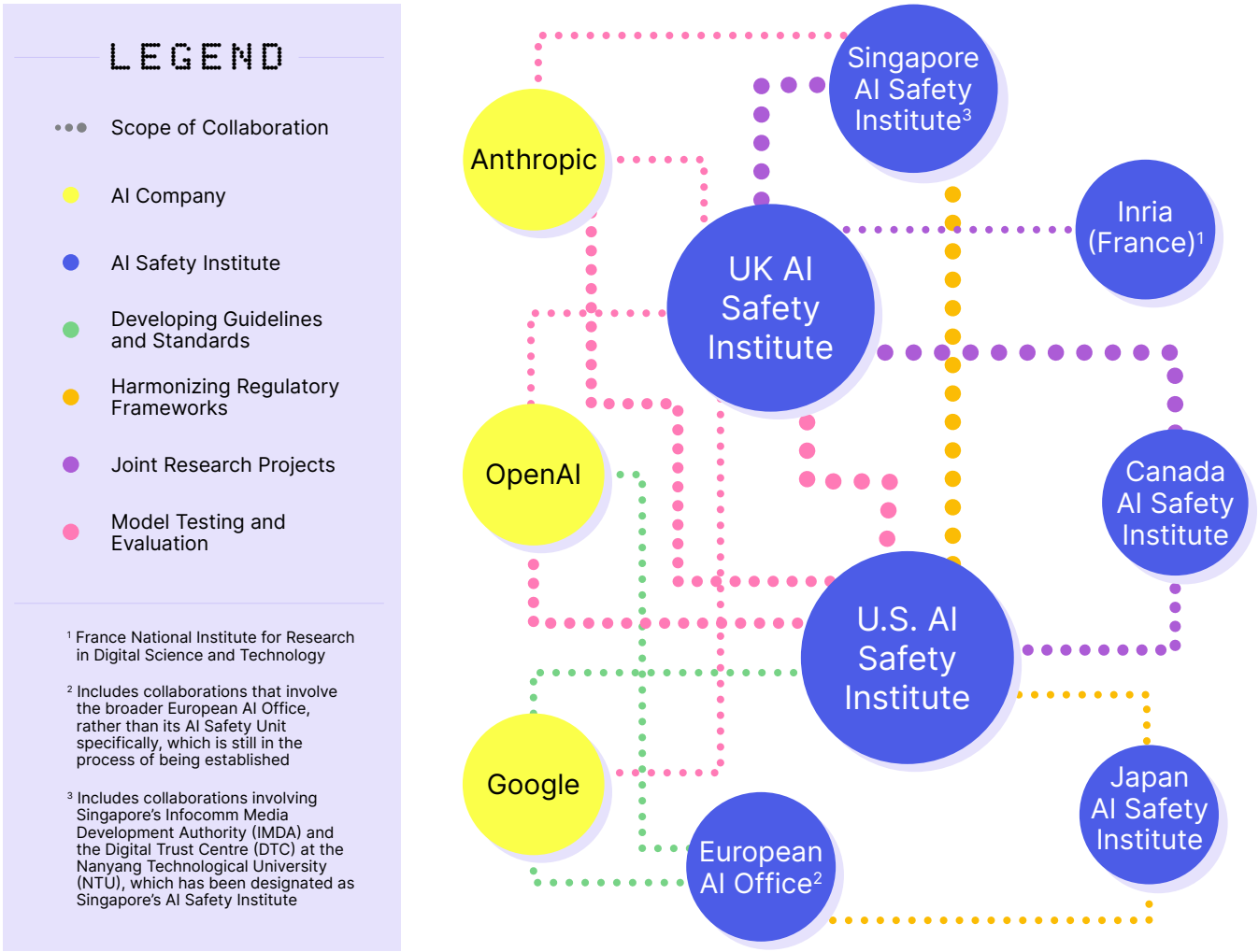
Contents

Executive Summary

AI Safety Institutes (AISIs) are specialized entities dedicated to ensuring the safe and ethical development and deployment of advanced AI systems. Addressing a broad spectrum of risks posed by these technologies, AISIs focus on advancing scientific and technical understanding, conducting rigorous evaluations of frontier models, shaping safety standards, and fostering global coordination on trustworthy AI.

Collaboration among AISIs is essential to amplify their collective capacity to mitigate AI risks effectively. By working together, they can enhance global standards and interoperability, streamline knowledge exchange, respond to incidents faster, and share resources and infrastructure. However, achieving this level of coordination is not without its challenges. Differences in national priorities, regulatory landscapes, resource availability, and the complexities of managing sensitive data create significant obstacles to seamless collaboration.

Despite these hurdles, AISIs have made meaningful progress in building partnerships, primarily through cross-sectoral collaborations and bilateral initiatives. A significant step



A network diagram of publicly announced bilateral collaborations between AISIs as well as with major AI companies.These collaborations are categorized into four primary areas: joint research projects, harmonizing regulatory frameworks, developing guidelines and standards, and model testing and evaluation. Line color indicates the primary area of collaboration, as inferred from the most emphasized topic in press releases about each partnership. Line size represents the scope of collaboration, with broader lines denoting a greater number of distinct areas of collaboration. This means that a larger dotted line represents that the entities collaborate in multiple areas, with the main form of collaboration indicated by the color of the line. A comprehensive source list for each connection in the diagram is available in Appendix I.

toward multilateral coordination was achieved with the launch of the International Network of AI Safety Institutes (AISI Network) in May 2024. The Network brings together members from diverse regions, including Australia, Canada, the European Union, France, Japan, Kenya, the Republic of Korea, Singapore, the United Kingdom, and the United States. Its inaugural formal gathering, scheduled for November 20-21, 2024, in San Francisco, aims to lay the groundwork for "advancing global collaboration and knowledge-sharing on AI safety."

> This report examines how the AISI Network can be structured and strategically integrated into the broader AI governance ecosystem to effectively fulfill its mandate toward secure, safe, and trustworthy AI.

Section 1 introduces AISIs, providing a high-level overview of their purpose and core functions. Section 2 examines the benefits and challenges of collaboration among AISIs while, in Section 3, the focus shifts to outlining the current state of collaboration. Building on this foundation, Section 4 considers potential structures of the AISI Network for enhanced coordination. Finally, Section 5 explores how AISIs can engage with and complement other prominent actors within the global AI governance ecosystem.

Together, these sections aim to provide an analysis of the current state of collaboration between AISIs and other key institutions, and propose strategic directions to deepen these collaborative efforts inside and outside of the Network, enhancing its impact in addressing global AI challenges.

How Could the AISI Network be Structured to Enhance Collaboration?

To strengthen collaboration among its members, the AISI Network could benefit from establishing a **central coordination arm**, or secretariat, to streamline and align efforts across institutes.

Responsibilities of the coordination arm could include aligning research agendas, facilitating working groups, standardizing evaluations, coordinating joint research programs, bridging research and policy, sharing technical expertise, managing member admission, drafting terms of reference, organizing meetings and events, securing network funding, and representing the Network internationally.

The AISI Network's San Francisco meeting presents a timely opportunity **to formalize the Network's coordination efforts by defining its scope, establishing clear criteria and terms for membership, and agreeing on concrete and actionable projects for collaboration.**

We propose three potential models for the AISI Network to consider:

Model	Benefits	Challenges
Rotating Secretariat: Member countries alternate as the Network's coordination arm, distributing administrative duties across nations (e.g. G7 model).	• **Shared Responsibility and Geographic Representation:** Ensures broad geographic representation and equitable influence in agenda-setting, preventing one member from dominating discussions. • **Diverse Agendas and Specialization:** Enables hosts to bring unique focuses and perspectives, potentially enriching the Network's expertise.	• **Shifting Priorities:** Allows rotating hosts to shift agendas, potentially leading to mission drift or mission creep. • **Continuity Issues:** Leads to inefficiencies, institutional "memory loss," and slower progress on long-term objectives due to frequent transitions.

Model	Benefits	Challenges
Static Secretariat in a Designated Country: A permanent administrative body located in a single country (e.g., World Bank model).	• **Institutional Stability and Consistency:** Facilitates long-term planning and maintains strategic continuity through a steady leadership structure. • **Established Infrastructure and Local Networks:** Utilizes established relationships and infrastructure to enhance administrative efficiency.	• **Perception of Bias:** Risk of a single host country dominating priorities, potentially alienating other members and deterring some regions from full participation.
Static Secretariat Hosted by an Intergovernmental Organization (IGO), such as the United Nations or the OECD: Embedded within the IGO for greater reach and legitimacy.	• **Global Credibility and Inclusiveness:** Facilitates coordination with countries lacking AISIs, and leverages the IGO's reputation for broad stakeholder engagement. • **AI Expertise and Established Networks:** Provides access to AI expert networks and multi-stakeholder collaboration channels, enhancing collective knowledge and capabilities. • **Established Diplomatic Safeguards:** Offers legal immunity for member institutes, ensuring secure and neutral platforms for international collaboration. • **Structured Funding:** Enables equitable funding mechanisms for sustainable support (e.g. proportional to GDP).	• **Bureaucratic Delays:** Complicates timely responsiveness to emerging AI risks due to the intricate processes of IGOs. • **Scope Limitations:** Risks reducing attention on frontier risks and limiting specialized knowledge for advanced AI governance in order to balance diverse national interests. • **Inefficient industry collaboration:** Slows industry engagement, making it more complex for AI labs to share resources or provide access to models, thereby restricting cooperation.

How Could the AISI Network Collaborate with Other Multilateral Efforts?

A well-coordinated AISI Network would have the capacity to engage and collaborate with key global actors, including the United Nations (UN) and its agencies, the Organisation for Economic Co-operation and Development (OECD), and other multilateral and regional organizations.

As specialized hubs, AISIs are uniquely positioned to contribute to and drive scientific consensus on advanced AI. By conducting targeted research and providing technical expertise, AISIs can complement the broader governance efforts of intergovernmental organizations and multilateral initiatives. This includes:

- Aligning technical evaluations and safety standards across borders.
- Systematically assessing the risks and benefits of emerging AI capabilities.
- Generating actionable insights for global policymakers.

Collaborating with the United Nations and its Agencies

We recommend that the AISI Network engage with UN processes selectively, prioritizing collaborations that preserve its independence and technical specialization. Similarly, to maintain inclusivity and fairness, the UN should balance its reliance on the AISI Network with contributions from member states and other stakeholders, ensuring that nations without dedicated AISIs are not marginalized.

Some collaborations between the AISI Network and UN could include ensuring that each member of the AISI Network actively participates in the UN International Scientific Panel, aligning the Summits and the Global Policy Dialogue on AI to build on each other's outcomes, facilitating provision of technical knowledge from the AISI Network to UNESCO for AI Readiness Assessments, and leveraging UN platforms to enhance the AISI Network's inclusivity and expand its reach.

Collaborating with the OECD

We recommend that the AISI Network harness the OECD's analytical expertise, extensive network, and proven ability to translate complex technical insights into actionable policy frameworks to collaboratively produce specific outputs, such as the *International Scientific Report on the Safety of Advanced AI*.

The AISI Network could play an active role in the newly formed partnership with the UN, by offering specialized expertise on frontier models and advanced AI risks. Additionally, it could also assist the OECD in monitoring the G7's Hiroshima Process AI Code of Conduct. Lastly, the AISI Network and the OECD could collaborate on the monitoring of AI incidents.

Expanding Collaboration with Other International Coalitions

To promote a truly global approach to AI safety, the AISI Network should also engage with other multilateral and regional efforts, **especially those that may lack the immediate capacity or intent to establish dedicated AISIs but still have a vested interest in the safe development of AI systems**.

For example, engaging with more non-Western entities like the China-BRICS Artificial Intelligence Development and Cooperation Center and regional organizations such as the African Union and ASEAN would infuse the AISI Network with necessary diverse perspectives. **The Network's inclusion of the European Union offers a model for engaging "regional AISIs,"** groups of countries collaborating to ensure representation in AI safety discussions without establishing their own domestic institutes. Early engagement with these entities will ensure that emerging AI governance frameworks are inclusive and resilient.

> By expanding its reach and building partnerships across diverse regions, the AISI Network can strengthen its impact, bring new perspectives to the table, and help shape a globally inclusive approach to AI safety and governance. These efforts would not only bolster trust and cooperation among stakeholders but also ensure that AI is developed and deployed safely and ethically on a global scale.

1. What Are AI Safety Institutes (AISIs)?

As the capabilities of artificial intelligence increase and its impact on society continues to grow, governments worldwide are establishing mechanisms to manage the risks and harness the benefits of these transformative technologies. Among the most prominent of these initiatives is the creation of AI Safety Institutes (AISIs)—specialized, government-funded entities tasked with overseeing the safe development of AI systems, especially those on the frontier of technological advancement.[1]

In November 2023, the United States became the first country to formally announce the establishment of an AISI, set to operate within the National Institute of Standards and Technology (NIST). The following day, against the backdrop of the inaugural AI Safety Summit at Bletchley Park, the United Kingdom followed suit, launching its own AISI as an evolution of its Frontier AI Taskforce and as part of the Department for Science, Innovation and Technology (DSIT).

In the following six months, Japan, the European Union (as part of the European AI Office), and Canada each launched their own AI Safety Institutes, followed by announcements from the Republic of Korea and Singapore, at the AI Seoul Summit in May 2024. Additionally, Australia, France, and Kenya revealed plans to launch AI Safety Institutes.

Together, these ten countries have joined forces to establish the International Network of AI Safety Institutes, which will convene its inaugural formal meeting in San Francisco in November 2024.

Despite notable differences among the AISIs, they generally operate as independent or semi-independent organizations with the primary goal of ensuring the safe development and deployment of advanced AI systems. Their core functions are multifaceted and include evaluating AI models for safety and alignment with regulatory standards, conducting risk assessments, and providing technical and scientific expertise to support AI policy development. Furthermore, AISIs are expected to play a critical role in testing and validating AI models before they reach the public, with a particular focus on models with high societal impact or those classified as "frontier AI."

By effectively carrying out these functions, AISIs have the potential to serve as essential safeguards against AI risks. Collaboration and coordination, both among AISIs and with other key actors in the AI governance ecosystem, are crucial to enabling them to achieve this role fully. In the next section, we explore the benefits of collaboration, as well as the challenges that have to be addressed.

AISIs have the potential to serve as essential safeguards against AI risks. Collaboration and coordination, both among AISIs and with other key actors in the AI governance ecosystem, are crucial to enabling them to achieve this role fully.

[1] While some AISIs are established as government entities, others may operate as independent or private institutions. This report does not presume the organizational structure or funding model of future AISIs.

2. Why Is AISI Collaboration and Coordination Important?

Benefits of Collaboration

Given the rapid evolution and expanding scope of AI technologies, there are several reasons why collaboration between AISIs is important:

Enhanced global standards and interoperability

Coordinated efforts among AISIs can help harmonize global standards and practices for AI safety, making it easier for AI models to be evaluated consistently across borders. Although each jurisdiction maintains the ability to establish regulatory requirements and obligations, this interoperability can streamline international regulatory compliance and facilitate smoother cross-border collaborations.

Streamlined knowledge exchange

Another benefit of AISI coordination is the establishment of accurate, timely exchanges of best practices, evaluation results, research findings, and other types of critical information (e.g. on AI incidents).

> Mechanisms such as joint testing exercises, personnel exchanges and secondments, training programs for technical capacity building, and collaborative reports can support this exchange, ultimately fostering a shared scientific understanding of AI safety.

Faster incident response and crisis management

A coordinated network of AISIs can also establish rapid communication channels and shared resources to address AI incidents or emerging threats. This improves the collective ability to mitigate risks and respond to crises effectively, which is increasingly important as the capabilities of AI systems continue to grow.

Resource and infrastructure sharing

Collaboration among AISIs can allow for the pooling of resources, such as specialized testing environments, advanced computational infrastructure, and unique datasets. This shared infrastructure may reduce costs and ensure that even smaller or less-resourced AISIs (or any other institutional equivalent of an AISI) have access to the tools necessary for rigorous AI safety evaluations.

Building technical capacity

Collaboration among AISIs can help mitigate disparities in technical capacity across nations. As publicly funded entities, AISIs already face challenges in competing with well-resourced private AI firms in securing a limited pool of top-tier AI engineers and scientists. By promoting exchanges of research and best practices, AISIs can alleviate some of this competitive pressure, ensuring that expertise and insights are shared across borders.

Coordinated initiatives—such as expert secondments, joint training programs, and shared technical resources—can create more equitable opportunities for participation in the AISI Network, enabling diverse nations to contribute valuable research and engage meaningfully in evaluating advanced AI systems.

Specialization

Coordination can provide opportunities for AISIs to specialize, whether focusing either on specific geographic regions for model testing or on particular thematic areas of AI safety. Thematic specialization coupled with robust knowledge-sharing, can allow individual AISIs to prioritize specific issues while benefiting from shared insights, reducing competition for limited talent and resources.

Mutual recognition of safety evaluations

To address the practical challenges of granting each AISI access to evaluate every frontier AI model, AISIs could establish a system for mutual verification of safety evaluations. This would allow companies to undergo evaluations through

their local AISIs, with resulting certifications acknowledged across the network as meeting shared safety standards, thus streamlining the evaluation process across borders and minimizing unnecessary duplication of efforts.

Challenges to Effective Collaboration

Notwithstanding its clear benefits, coordination amongst AISIs is not easy. It also presents a range of challenges, many of which stem from the complexities of managing diverse national interests and priorities, geopolitical tensions, legal frameworks, and funding disparities.

National security and competitiveness concerns

While information sharing promotes collaboration and trust, it also raises significant national security and competitiveness concerns for governments.

> As AI becomes increasingly integral to national security, governments are likely to be cautious about the unintended disclosure of classified or proprietary information, particularly in relation to AI models with military or strategic applications.

The dual-use nature of AI compounds these concerns, as even commercial AI advancements could indirectly enhance a rival nation's military capabilities.

Governments are also likely to be reluctant to share data that could reveal proprietary insights about their domestic companies' AI models, as this could lead to a competitive disadvantage. Such concerns are particularly acute when dealing with geopolitical rivals, where shared information might empower foreign industries or accelerate technological advancements in competing countries. This hesitancy extends to key details about model architecture and evaluation techniques, which, if disclosed or inadvertently leaked by allies with weaker cybersecurity controls, could diminish the market position of leading companies or even compromise a nation's strategic edge in AI innovation.

Unless AISIs opt to withhold all potentially sensitive information—a choice that could significantly undermine the Network's knowledge-sharing goals—establishing robust information-sharing protocols is essential to mitigate the risks of data leakage. These protocols should enforce stringent data security measures, safeguarding against unauthorized access and minimizing the chances of inadvertent information exposure.

Disruption as a result of shifts in national agendas and/or elections

Political shifts, particularly election outcomes, can disrupt the stability and collaborative focus of each AISI as well as the Network as a whole. For instance, with Donald Trump's reelection, concerns have arisen over whether the United States will sustain its current role in international AI governance or pivot to a less cooperative approach, with even less governance. The potential repeal of the Biden Administration's Executive Order 14110 and intensified U.S.-China competition may deprioritize global collaboration in favor of rapid AI development with minimal safety oversight. Such changes can result in a shift in scope for the U.S. AISI.

Data sharing and access issues

There may also be numerous legal and regulatory obstacles to some forms of data exchange among AISIs. Different countries have varying regulations and policies regarding data privacy, intellectual property rights, and access to proprietary datasets. For example, data protection regulations like the GDPR in Europe and a patchwork of federal and state privacy laws in the United States create inconsistent standards for information sharing, potentially restricting the flow of critical safety evaluation data. AISIs may be reluctant to engage fully in extraterritorial international collaborations if their partners are unwilling or unable to share crucial data, hindering progress in global AI safety efforts.

Funding disparities

Unequal financial commitments to AISIs could cause friction. For example, the UK has currently pledged significantly more funding for its AISI compared to other countries, potentially

leading to concerns about disproportionate contributions to global AI safety efforts. This could lead to a situation where some countries have greater influence, or where other countries benefit more from collaboration without offering comparable resources, similar to issues seen in organizations like NATO. At the same time, insufficient funding could threaten to compromise the effectiveness of the AISI efforts.

Competition for talent and resources (e.g. compute)

AISIs are often publicly funded and may struggle to compete with well-funded private sector players for top AI talent and resources, such as GPU chips. These funding disparities and the limited pool of highly specialized AI experts could introduce competitive pressures among AISIs, adding strain to collaborative efforts and potentially limiting the network's ability to advance shared objectives.

Redundancy with other multilateral efforts

AISIs could risk becoming redundant or obstructive to existing multilateral efforts or organizations such as international standards setting bodies. These organizations are typically already equipped with legitimacy and resources to set standards or lead intergovernmental research projects. This overlap could lead to a diffusion of responsibility and unnecessary bureaucracy, ultimately undermining the effectiveness and mission of AISIs.

Complexity in specialization

Allowing each AISI to specialize in a specific area of expertise within a global network will help to reduce redundancies, as mentioned previously, but could also add complexity to the ecosystem. It could result in overreliance between partner AISIs in critical safety areas that could be highly costly if, for instance due to political shifts or diplomatic tensions, AISIs decide to stop collaborating with each other.

Varying legal mandates

An absence of a unified international AI regulatory structure, and differences in legal mandates across AISIs introduce operational barriers to effective coordination.[2] For instance, the European AI Office, under the EU AI Act, has statutory powers to impose regulations on AI companies and enforce compliance, whereas the U.S. and UK AISIs operate on a basis of voluntary collaboration without explicit regulatory authority. This disparity is evident in the EU's ability to require mandatory testing and documentation of frontier AI models, while other AISIs must individually negotiate access with AI developers.[3]

Navigating bureaucratic hurdles

A final challenge is that formal coordination structures, such as the AISI Network, can slow down activity, adding bureaucratic complexity and delaying decision-making. This reduces the agility and flexibility of informal partnerships, which could be more effective at quickly responding to emerging AI risks. Balancing structured coordination with the need for nimble, adaptive responses remains a persistent challenge.

AISIs may be reluctant to engage fully in extraterritorial international collaborations if their partners are unwilling or unable to share crucial data, hindering progress in global AI safety efforts.

[2] Even with the emergence of such a structure or framework, historical precedent suggests that nations are likely to maintain selective engagement—adopting and implementing provisions that align with their domestic priorities and capabilities while opting out of others, creating a de facto "à la carte" system of cooperation.

[3] As a further example, intellectual property and copyright considerations for training foundation models are expected to spark intense debate in the coming years, with countries and regions potentially diverging significantly in their approaches to regulating data privacy and ownership.

3. How Are AISIs Collaborating Currently?

Even before the formal establishment of the AISI Network, AISIs had already initiated collaborations, both with one another and with non-governmental actors such as AI companies and academia. This section provides an overview of the current state of collaboration between AISIs across three levels: cross-sectoral, bilateral, and multilateral.

Cross-Sectoral Collaborations

Cross-sectoral collaboration involves structured partnerships between AISIs and non-governmental entities—including industry leaders, academic institutions, and civil society organizations—to advance comprehensive safety evaluation and standards development.

AISIs are increasingly partnering directly with leading AI companies to conduct joint evaluations, red-teaming exercises, and safety testing:

- The UK AISI has established partnerships with top research organizations, securing privileged access to cutting-edge AI models from leading companies to test for cyber, chemical, biological, and agentic capabilities.
- Similarly, the U.S. AISI has signed a Memorandum of Understanding with Anthropic and OpenAI, enabling formal collaboration on AI safety research, testing, and evaluation. This agreement provides the U.S. AISI access to these companies' newest models both before and after their public release.
- The Biden Administration's October 2024 U.S. Memorandum on Advancing the United States' Leadership in Artificial Intelligence designates the U.S. AISI as the lead point of contact between private sector AI developers and government to facilitate voluntary pre- and post-deployment safety testing of frontier AI models.
- In Singapore, the Infocomm Media and Development Authority (IMDA) and the AI Verify Foundation have partnered with Anthropic to conduct red-teaming exercises across languages and cultural contexts.

Public-private partnerships can also facilitate access to compute resources and datasets for AISIs, supporting their efforts to develop new state-of-the-art testing and evaluation techniques:

- The UK government has prioritized access to over £1.5 billion in computational resources through its AI Research Resource and exascale supercomputing program.
- Singapore's AI Verify Foundation brings together stakeholders from various sectors, including industry, academia, and government, to collaborate on building testing infrastructure. It recently launched the Project Moonshot platform, which combines various datasets to create customizable testing packages tailored to specific needs

Collaboration with academia and industry plays an important role in developing policies, facilitating compliance, and raising awareness of AI safety issues:

- The U.S. AI Safety Institute Consortium brings together over 280 organizations, including AI creators, users, academics, and civil society, to develop empirically backed guidelines and standards for AI policy.
- Similarly, Japan's AISI is working with academic and industry partners to promote AI safety through seminars and educational materials.
- The European AI Office has also been actively engaging a variety of stakeholders, although this has not specifically been carried out by its AI safety unit, which is still being established. Industry groups, academia, and civil society from around the world are involved in the drafting of the Code of Practice,

with the working group Chairs[4] coming mainly from academia and civil society. Additionally, the AI Office also oversees the AI Pact, encouraging early compliance with the forthcoming AI Act from AI model providers.

Bilateral Collaborations

Bilateral collaboration between AISIs refers to formal partnerships between two AISIs that enable structured exchange of technical expertise, testing methodologies, and safety evaluation results while maintaining institutional independence. While multilateral networks offer broader perspective diversity and resource pooling, bilateral partnerships often enable more rapid progress on targeted objectives due to reduced coordination overhead and simplified decision-making processes. However, this efficiency advantage varies based on factors like partnership scope, existing institutional relationships, and alignment of safety evaluation methodologies.

Joint model testing exercises and research programs allow AISIs from different countries to combine their expertise and share specialist knowledge:

- In April 2024, the UK and U.S. AISIs signed a Memorandum of Understanding (MoU) for working together to develop tests for the most powerful AI models. They intend to perform at least one joint testing exercise on a publicly accessible model.

Personnel exchanges between AISIs can support deeper collaboration and knowledge transfer:

- The MoU between the UK and U.S. AISIs aims to facilitate personnel exchanges, while similar commitments have been made between the UK and Canada to promote professional development and collaboration. The United States and Singapore also plan on launching an "AI Talent-Bridge" program".

Several joint research programs between institutes have also been announced:

- The UK and Canada commit to creating pathways for the sharing of expertise to bolster existing testing and evaluations work and to jointly identify other priority areas for research collaboration. Notably, the UK AISI will share its allocation of priority access to the UK AI Research Resource with the Canadian AISI on their joint research.
- Along with the U.S. AISI, the UK and Canada are also reportedly planning on launching a program of research to catalyze the field of "Systemic AI Safety," which refers to the safeguarding of societal systems into which AI is being deployed.
- In June 2024, the U.S. AISI announced collaboration with the Singapore AISI on "advancing the science of AI Safety."
- The UK and France committed £800,000 in new funding to deepen research and AI links through programs like Horizon Europe.
- In November 2024, the UK and Singapore AISIs announced a partnership focused on collaborating closely to advance research and develop a shared framework of policies, standards, and guidelines.

Bilateral regulatory "crosswalks" can harmonize frameworks and improve interoperability, while regular dialogues between countries build a deeper understanding of each other's needs and objectives:

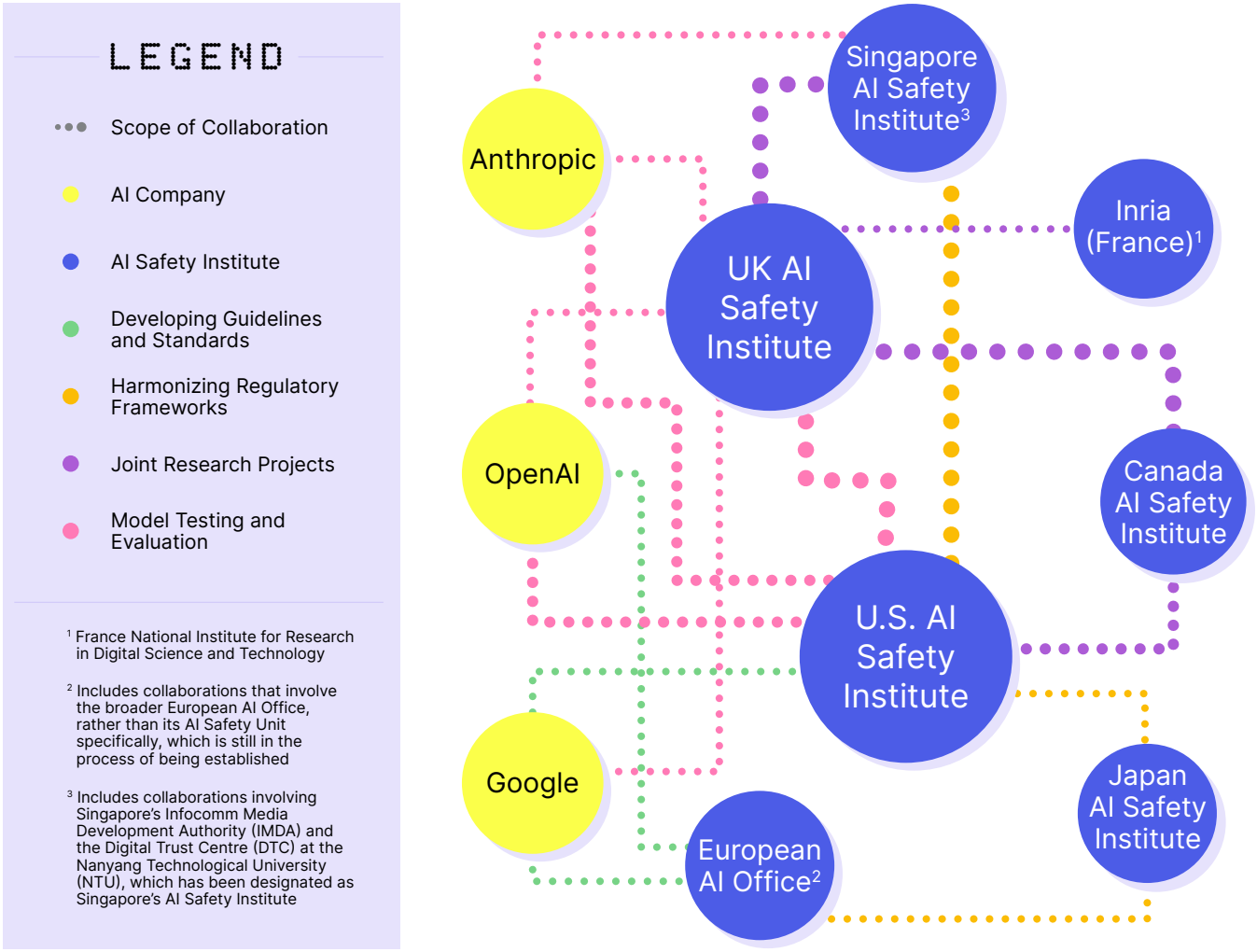
- The crosswalks between Japan's AI Business Guidelines and the U.S. NIST AI Risk Management Framework (RMF), are helping to harmonize AI safety frameworks and ensure mutual understanding of each country's objectives and regulatory needs.
- The United States and Singapore AISIs announced their intention to map their respective frameworks for generative AI, exploring collaboration on testing, guidelines, and benchmarks.

[4] The geographic distribution of Chairs and Vice-Chairs, including two from North America, suggests that the development of the Code of Practice is a more globally integrated process (albeit still Western) than an entirely Europe-based endeavor.

- Japan and the EU have pledged to enhance collaboration between their respective AI offices, aligning their efforts with global initiatives like the G7 Hiroshima AI Process and the AI Pact.

The below network diagram illustrates the extent of current and planned bilateral collaborations between AISIs and the three major AI companies—Google, OpenAI, and Anthropic—that have publicly engaged most extensively with AISIs. These collaborations are categorized into four primary areas: joint research projects, harmonizing regulatory frameworks, developing guidelines and standards, and model testing and evaluation, as indicated by the color-coded lines in the legend. The width of each line represents the "scope" of collaboration, which reflects the number of distinct areas in which each entity pair is actively cooperating.

The diagram highlights Singapore, the UK, and the U.S. AISIs as central hubs in these bilateral partnerships. The UK and U.S. AISIs, in particular, demonstrate extensive engagement across multiple areas with a broad range of partners, including all three major AI companies.



A network diagram of publicly announced bilateral collaborations between AISIs as well as with major AI companies.These collaborations are categorized into four primary areas: joint research projects, harmonizing regulatory frameworks, developing guidelines and standards, and model testing and evaluation. Line color indicates the "main" area of collaboration, generally inferred from the most emphasized topic in press releases about each partnership. Line size represents the "scope" of collaboration, with broader lines denoting a greater number of distinct areas of collaboration. This means that a larger dotted line represents that the entities collaborate in multiple areas, with the main form of collaboration indicated by the color of the line. A comprehensive source list for each connection in the diagram is available in Appendix I.

Multilateral Collaboration: The AI Safety Institute Network

The key fora for multilateral partnership among AISIs at present is the International Network of AI Safety Institutes. This network includes members from Australia, Canada, the European Union, France, Japan, Kenya, Republic of Korea, Singapore, the United Kingdom, and the United States.

Launched in May 2024 at the Seoul AI Summit, the AISI Network's stated mission is "to promote the safe, secure, and trustworthy development of AI." The first formal gathering of technical AI experts from these member countries and the EU is scheduled for November 20-21, 2024, in San Francisco. This inaugural meeting will focus on aligning priority work areas for the Network and establishing a foundation for "advancing global collaboration and knowledge sharing on AI safety".

The following sections examine how this vision of global collaboration could take shape and explore how the AISI Network can be structured and strategically positioned within the broader AI governance ecosystem to fulfill its objectives effectively.

> *Bilateral collaboration between AISIs ... enables structured exchange of technical expertise, testing methodologies, and safety evaluation results while maintaining institutional independence. While multilateral networks offer broader perspective diversity and resource pooling, bilateral partnerships often enable more rapid progress on targeted objectives due to reduced coordination overhead and simplified decision-making processes.*

4. How Could the AISI Network Be Structured to Enhance Collaboration?

Effective internal coordination is crucial for the AISI Network to engage meaningfully with external actors in global AI governance. A well-organized network would not only amplify AISIs' collective influence but also optimize resource use, strengthen AI safety protocols, and present a unified voice in global discussions.

The November 2024 meeting in San Francisco presents an important opportunity to formalize the Network's structure and reinforce members' commitment to a set of shared objectives. This gathering should **prioritize defining mechanisms for collaboration—such as leadership frameworks, Terms of Reference (TORs) for members, mechanisms for secure information exchange, shared digital platforms, and regular convenings**—that will help the Network function as a unified, strategic force in advancing AI safety.

Below, we outline approaches to cultivating unity among members and propose organizational structures to support these goals, thereby enhancing both internal alignment and external impact.

Establishing a Central Coordination Arm

To enable internal coordination and collaboration, the AISI Network would benefit from a central coordination arm, or secretariat, dedicated to aligning member efforts and addressing global AI safety challenges cohesively. This secretariat would function as the Network's core structure, responsible for setting objectives, synchronizing research initiatives, standardizing methodologies, and handling external communications.

> By centralizing these tasks, a dedicated secretariat could alleviate the burden on individual AISIs to sustain joint initiatives informally, ensuring continuity across projects and regular network activities.

To date, the few secretariat-like functions that have been carried out within the AISI Network have primarily been managed by individual member AISIs, with the well-resourced UK AISI playing a central coordinating role across several initiatives and the U.S. Departments of State and Commerce taking the lead in organizing the upcoming San Francisco meeting. However, if this trend were to continue, the Network risks becoming dominated by specific national interests. Establishing a dedicated, neutral secretariat could mitigate this imbalance, promoting a more inclusive and balanced approach to the Network's activities and strategic direction.

When considering how best to structure the AISI Network's coordination arm, valuable lessons can be drawn from the experience of the Global Partnership on AI (GPAI). In GPAI's case, centralizing much of the substantive work within Canada and France made it challenging for other members to engage fully and limited meaningful contributions from newer participants. This perception of concentrated influence contributed to the integration of GPAI within the OECD in 2024. To avoid similar challenges, the AISI Network might benefit from a coordination arm that ensures transparent roles and accessible pathways for all members to contribute effectively.

Responsibilities of the coordination arm could include:

- **Aligning Research Agendas:** Coordinating research goals across AISIs to prevent duplication, optimize resource use, and ensure efficient project execution.
- **Facilitating Working Groups:** Establishing and managing working groups focused on priority AI safety topics.
- **Standardizing Evaluations:** Coordinating standards, evaluations, and information-sharing protocols across members.
- **Coordinating Joint Research Programs:** Enabling AISIs to collaborate on shared research initiatives, pooling expertise on critical AI safety issues.

- **Bridging Research and Policy:** Acting as a liaison between technical research bodies and policymakers, offering evidence-based recommendations to shape governance frameworks.
- **Sharing Technical Expertise:** Facilitating access to technical resources for academic institutions, policymakers, and government officials to support their AI risk mitigation efforts.
- **Managing Member Admission:** Developing a formal process for admitting new members.
- **Drafting Terms of Reference:** Establishing a Memorandum of Understanding and Terms of Reference for all members, with mechanisms to monitor adherence.
- **Organizing Meetings and Events:** Scheduling and planning network events, including member convenings and public engagements.
- **Securing Network Funding:** Arranging funding sources and overseeing the collection of contributions from members.
- **Representing the Network Internationally:** Acting as the primary representative for the Network at external multilateral forums.

Absent a central coordinating arm, many of the above responsibilities would likely remain decentralized, managed individually by member AISIs or on a bilateral basis:

- Each AISI would likely independently set its research agenda without a unified framework. At best, coordination of joint programs would be fragmented, with individual AISIs forming bilateral or ad-hoc partnerships rather than leveraging collective resources.
- Member AISIs could end up setting their own evaluation standards and methodologies, leading to potential inconsistencies in safety assessments across the network.
- Each AISI would engage in external communications and representation independently, potentially resulting in a lack of unified messaging and less cohesive engagement with external stakeholders.

To structure the secretariat effectively, we explore three potential models, each with unique strengths and challenges that have significant implications for the Network's cohesion, direction, and operational impact.

Rotating Secretariat

> **The rotating secretariat model would involve member countries taking turns serving as the Network's coordination arm.** By rotating these duties, the model promotes inclusivity and distributes the logistical and financial responsibilities associated with guiding the Network, rather than placing this burden on a single country year after year.

This approach mirrors the G7's rotation presidency, where each member nation takes turns hosting and setting the agenda for annual meetings.

For example, the upcoming event in San Francisco will be hosted and organized by the U.S. government, while future AISI Network meetings could rotate to other member countries. This rotation reflects the precedent set by the series of global AI summits, which began as the AI Safety Summit at Bletchley Park in the UK in November 2023, continued with the AI Seoul Summit in May 2024, and will move to France for the AI Action Summit in February 2025.

BENEFITS:

Shared responsibility and geographic representation: This decentralization prevents any one member from dominating discussions or priorities, cultivating a sense of shared responsibility and equitable influence over agenda-setting. It also allows for broader geographic representation and ensures that activities reflect varied perspectives on AI safety.

Diverse agendas and specialization: Each host nation can introduce specific topics based on their national interests or AI advancements, enriching the Network's collective knowledge, and allowing the entire network to benefit and learn from the specialization of particular countries' AISIs. For instance, AISIs within regions with strong regulatory frameworks, such the EU, may focus on policy best practices, while those with advanced technical research institutions, such as the U.S. or UK, may emphasize technological safety measures. Countries that excel at fostering public-private partnership, such as Singapore, could lead in this area.

CHALLENGES:

Shifting priorities: A rotating secretariat model introduces the risk of fluctuating agendas and goals with each new host country, potentially disrupting continuity in focus. Each host may emphasize different facets of AI safety based on its own priorities, which can dilute the core objectives over time.

For instance, the inaugural AI Safety Summit at Bletchley Park in the UK in November 2023 was specifically focused on frontier AI risks. However, the subsequent "AI Seoul Summit" in May 2024 broadened the agenda, dropping the explicit "safety" focus from the name. Now, the upcoming "AI Action Summit" in France in February 2025 is expected to cover frontier risks as just one of five themes, alongside accessibility, innovation, and AI's impact on the labor market. Although this broadening of focus widens the spectrum of risks addressed, it diverts attention from the original intent, demonstrating how a rotating model can lead to mission drift as each host brings varying priorities to the table.

Continuity issues: The rotating model may also cause logistical inefficiencies and institutional "memory loss", as each new host must, to a certain extent, re-establish administrative capacity and rebuild key relationships. Frequent leadership transitions could therefore disrupt continuity, delay important decisions, and slow progress on long-term objectives.

Static Secretariat in a Designated Country

The static secretariat model would establish a permanent administrative body within a designated host country, with dedicated staff, standardized protocols, and stable funding mechanisms. This model enables long-term institutional development while requiring careful consideration of governance structures to ensure equitable representation. This setup mirrors traditional international organizations like the World Bank or the International Monetary Fund, both headquartered in Washington, D.C. While the secretariat would have a fixed location, a rotating chair or presidency within the structure could ensure balanced representation and responsiveness to all members. The secretariat would provide continuous administration, coordination, and support regardless of changes in leadership.

BENEFITS:

Institutional stability and consistency: A permanent secretariat enables a stable, cohesive agenda, allowing for more effective pursuit of long-term goals, as well as multi-year strategy including research programs and policy initiatives. With consistent leadership, the secretariat can set both short- and long-term priorities with minimal disruption when the presidency shifts to a different country. This stability supports sustained progress and avoids the shifting priorities that can accompany rotating leadership. This structure also helps retain institutional knowledge over time, preventing the "memory loss" that can occur when responsibilities are frequently transferred between hosts.

Established infrastructure and local network: A permanent headquarters enables ongoing administrative efficiencies, as systems, staff, and relationships can be built and refined over time. This continuity reduces the time and costs associated with transferring responsibilities to new hosts and minimizes the learning curve for leadership and staff.

For example, a U.S.-based secretariat could capitalize on its proximity to and influence over major AI companies to facilitate industry engagement and technical collaboration. A Singapore-based secretariat could offer a

balanced location that bridges both Western and Asian AI markets. A UK-secretariat could leverage the UK AISI's substantial funding to ensure a strong infrastructure. A secretariat at a neutral location such as Switzerland could be a strategic and impartial choice appealing to a broad range of stakeholders.

CHALLENGES:

> **Perception of bias:** Choosing a host country may prove contentious, as establishing a permanent headquarters in one country could engender concerns about neutrality, and be perceived as favoring the interests of the host country's companies or regulatory standards.

For example, while having the AISI Network headquartered in the United States would offer logistical advantages due to its proximity to the majority of frontier AI companies, some countries may hesitate to endorse this location. Non-U.S. stakeholders may view a U.S.-based secretariat as too closely tied to U.S. priorities, particularly given that the AISI is affiliated with NIST, which operates under the Department of Commerce–a department focused on furthering U.S. economic interests.

Conversely, hosting the AISI Network in Europe could raise concerns for others, as it would place the secretariat within the regulatory landscape shaped by the EU AI Act, which emphasizes stringent oversight of advanced AI models. This close proximity to the European AI Office could create apprehension that the Network's objectives may be swayed toward the EU's regulatory approach, potentially diverging from other countries' more laissez-faire policy preferences.

Static Secretariat Hosted by an Intergovernmental Organization

Rather than building a new entity, the AISI Network could choose to embed its secretariat within an existing intergovernmental body, such as the UN or the OECD. Leveraging an established organization would likely provide built-in diplomatic channels, administrative support, and international credibility. A similar precedent exists in the OECD's support for the Global Partnership on Artificial Intelligence (GPAI), where a dedicated secretariat facilitated coordination and provided administrative support.[5]

BENEFITS:

Global credibility and inclusiveness: Establishing the secretariat within a respected intergovernmental organization would enhance its credibility and make it easier to engage a broad range of stakeholders, including academia, civil society, and countries that do not have AISIs.

AI expertise and established networks: Intergovernmental organizations have robust, longstanding networks and working groups focused on AI safety and governance, which the AISI Network could tap into immediately. This access allows the Network to draw on diverse perspectives and benefit from established expertise in governance and safety standards. For example, the OECD's AI Policy Observatory offers a rich resource for international collaboration on AI policy, while the UN's High-Level Advisory Body on AI Governance brings together experts from various sectors to advise on global AI governance issues.

Structured funding: Many intergovernmental organizations use tiered funding models (e.g, based on GDP contributions), ensuring that the financial burden is shared equitably. This system can prevent wealthier nations from exerting undue influence over the network's objectives and agenda. The organized funding structure of international organizations offers both stability and equity.

For example, the OECD's tiered contribution model, based on member states' GDP, ensures sustainable funding while preventing financial leverage from dominating the agenda. However, this must be balanced against the need for rapid resource deployment in response to emerging AI safety challenges.

[5] Recently, GPAI fully merged into the OECD, further integrating its activities within the organization's established infrastructure and networks.

Established diplomatic safeguards: Operating under an inter-governmental body affords certain diplomatic and legal protections, allowing for a more secure environment for sensitive discussions. IGOs typically have legal immunity under international law, which protects them from lawsuits and legal actions in most jurisdictions. These protections could enhance trust among members, knowing that exchanges are safeguarded from unilateral national pressures, and would allow the Network to carry out its missions without the threat of constant litigation.

CHALLENGES

Administrative bureaucracy: Intergovernmental organizations are often characterized by complex bureaucratic processes, which could delay urgent responses to rapidly evolving AI safety threats. Integrating the AISI Network within a larger intergovernmental organization may reduce its agility, as lengthy approval cycles, consensus-building, and interdepartmental coordination can slow decision-making. For instance, the United Nations' periodic report-generation process, as seen with the Intergovernmental Panel on Climate Change (IPCC), typically spans five to eight years, making it challenging to keep pace with the fast-moving AI landscape. This administrative inertia could hinder timely assessments and actionable insights into pressing AI safety issues.

Scope limitations: A UN-driven approach would need to balance member states' varying interests in AI risks versus economic and strategic benefits, potentially creating tensions in the secretariat's scope. For example, limiting the scope to advanced AI risks might alienate countries prioritizing the economic potential of AI, while a broader focus may dilute the critical examination of frontier AI models. To engage a diverse membership, the AISI Network within the UN would likely need to create multiple working groups, each focusing on specific issues such as economic impacts, ethical considerations, and technical risks, but this would further complicate coordination and speed of outputs.

Inefficient industry collaboration: Placing the AISI Network under the management of an intergovernmental organization (IGO) could present challenges in engaging with the private sector, where much of the advanced AI research and proprietary data resides. The institutional red tape inherent in a multilateral framework might slow collaboration, particularly on matters involving sensitive or national security-related information. AI labs may hesitate to grant AISIs access to their models unless the process for doing so is clearly defined and managed effectively by the IGO. Similarly, member countries could be reluctant to share critical proprietary data, citing confidentiality concerns and competing national interests. This reluctance could be greater compared to the trust and flexibility offered by a smaller, more closely aligned network of allied nations.

Defining the Network's Scope, Membership Criteria, and Concrete Projects for Collaboration

A key deliverable for the upcoming AISI Network could be the adoption of Memoranda of Understanding (MoUs) that clearly outline the Network's mission, objectives, and overall scope. These documents should expand on the principles outlined in the Seoul Statement and not only provide a cohesive framework for collaboration between network members but also help external stakeholders recognize the unique contributions and value the AISI Network brings to global AI safety and governance efforts.

The AISI Network will also need to establish clear membership procedures and transparent terms for current and prospective members. To address the potential complexities of expansion, the members might also consider adopting a tiered membership structure, allowing various stakeholders to participate at different levels of commitment and information access. This approach would mirror models used by the Financial Action Task Force (FATF), and the OECD, which includes core members alongside "Key Partners."

Membership Terms of Reference (TORs) could outline expectations for active participation, research contributions, and adherence to shared principles of transparency and safety. TORs should also establish fair funding standards to prevent wealthier nations from disproportionately influencing the Network's agenda. These terms must ensure inclusivity, allowing resource-constrained nations to participate meaningfully while avoiding exclusions based on geopolitical tensions.

In addition, the AISI Network could prioritize a few strategically chosen, high-impact projects that are both feasible and achievable for the members to collaborate on in the near term. This approach will ensure sustained momentum following the AI Seoul Summit, while also demonstrating tangible progress ahead of the AI Action Summit in Paris, in February 2025.

In the short to medium term, priority projects could include:

- establishing consensus on safety standards,
- collaborating to improve evaluation methodologies,
- defining which types of information should be shared between AISIs,
- and developing robust mechanisms to facilitate these exchanges.

Over the medium term, the Network should develop a defined **research agenda** centered on technical safety priorities. This could include:

- Creating a **common glossary** of technical terms.
- Developing a unified, evidence-based approach to **testing and evaluation methodologies** to streamline cross-institute collaboration.

Such foundational work would not only enable more ambitious joint initiatives but also position the AISI Network as a global leader in advancing AI safety and governance.

Effective internal coordination is crucial for the AISI Network to engage meaningfully with external actors in global AI governance. A well-organized network would not only amplify AISIs' collective influence but also optimize resource use, strengthen AI safety protocols, and present a unified voice in global discussions.

5. How Could the AISI Network Collaborate with Other Multilateral Efforts?

An effectively coordinated AISI Network would be well-positioned to engage with prominent global actors such as the United Nations (UN) and its specialized agencies, the Organisation for Economic Co-operation and Development (OECD), as well as other multilateral and regional organizations. These partnerships would play a crucial role in global efforts toward AI safety and governance.

AISIs are emerging as pivotal actors within the global AI safety ecosystem, but their long-term effectiveness will hinge on their ability to strategically complement and integrate with established organizations.

Without clear coordination, there is a significant risk of fragmented efforts, resource inefficiencies, and competing priorities among parallel initiatives—dynamics that could dilute collective impact and hinder progress toward shared AI safety goals.

AISIs can serve as specialized hubs, offering targeted research and technical expertise on advanced AI to help build and drive scientific consensus.

> Their insights could be instrumental for enriching the policy-making and standard-setting efforts of global AI governance institutions, providing the nuanced, technical input necessary for robust AI safety frameworks. Additionally, AISIs can serve as vital intermediaries, bridging the gap between the fast-paced developments of the AI industry and the more deliberate processes of international regulatory bodies.

Specifically, AISIs can help align technical evaluations and safety standards across borders, clarify the risks and benefits of emerging AI capabilities, and generate actionable insights for global policymakers.

Below we outline specific opportunities for collaboration between AISIs and key global actors such as the UN and the OECD, as well as potential avenues to expand partnerships beyond these organizations.

Collaborating with the United Nations and its Agencies

The UN's global reach and role in multilateral governance make it a natural partner for AI safety initiatives. The UN's involvement in AI governance has recently gained momentum, with the High-Level Advisory Body (HLAB) publishing its final report in 2024 and the adoption of the Pact for the Future at the Summit of the Future including the Global Digital Compact (GDC). Both of these initiatives endorsed recommendations to establish a multidisciplinary International Scientific Panel on AI and initiate a Global Dialogue on AI Governance.

For the AISI Network to solidify its position in the ecosystem, meaningful collaboration with the UN will be essential, especially if the UN goes on to occupy a broader convening role within the global AI governance regime complex, as many commentators and governments have been calling for.

Mandated with a specialized focus on advanced AI, the AISI Network could offer agile, timely insights to the UN's broader assessments and reports, enriching the global discourse on AI safety. This synergy would pair the UN's inclusivity with AISI's technical rigor, enabling a more holistic approach to managing AI risks. For instance, the UK AISI's international scientific report could complement UN initiatives, such as the UNHLAB's Governing AI for Humanity. The UN reports can bring together global policymakers and provide an inclusive, high-level overview, while the AISI report can offer a more focused, in-depth analysis of AI safety challenges from a technical perspective.

To preserve the unique strengths of both

entities, however, i**t is recommended that the AISI Network engage with UN processes selectively, focusing on collaborations that do not compromise its independence and technical specialization**. Over-integration could dilute the AISI Network's impartial focus on risks from frontier AI, and aligning its outputs too closely to the UN's multilateral processes might slow down reporting, add pressure to harmonize findings, and reduce the autonomy of AISI's assessments.

Likewise, the UN should avoid an overreliance on the AISI Network as its sole source for analyzing advanced AI risks. Relying too heavily on AISI input could lead some member states to feel marginalized, which may undermine their support for UN-led AI initiatives. Instead, the UN should continue to integrate a diverse range of international perspectives, including input from countries without AISIs and alternative international AI governance networks, such as the China-BRICS Artificial Intelligence Development and Cooperation Center.

With these concerns in mind, potential areas of collaboration between the AISI Network and the UN might include:

Engaging with the UN International Scientific Panel

Each AISI Network member should designate a representative to actively participate in the newly announced UN International Scientific Panel. The Network should collectively monitor and contribute to the UN International Scientific Panel development, ensuring it reflects the technical expertise and collaborative research produced by both individual AISIs and the Network as a whole. As UN member states negotiate the panel's modalities, AISIs can play a pivotal role by working closely with their government counterparts to align priorities and ensure their perspectives are well-represented. Early involvement from AISIs could not only enhance the panel's alignment with the Network's insights but also build greater buy-in from member states, increasing the likelihood of the panel's broad endorsement and successful implementation.

Aligning the Summits and the Global Policy Dialogue on AI

Despite being distinct institutions, AISIs and the Global Summits are closely interconnected, with many viewing the Summits as critical platforms for AISIs to announce initiatives, make commitments, and shape the global agenda. To maximize impact, it is crucial for the outcomes of each Summit to feed directly into the UN's forthcoming Global Policy Dialogue on AI, creating a continuous and iterative process. Rather than restarting conversations at each convening, these dialogues should build upon prior insights, agreements, and recommendations, fostering a cohesive and evolving approach to AI governance. Additionally, to ensure legitimacy and inclusivity, these discussions must prioritize the participation of representatives from the Global Majority and other underrepresented stakeholder groups, amplifying diverse perspectives in shaping the global AI landscape.

Providing technical knowledge for UNESCO's AI Readiness Assessments

The AISI Network, as a central hub for AI safety best practices, could assist UNESCO

> *For the AISI Network to solidify its position in the ecosystem, meaningful collaboration with the UN will be essential, especially if the UN goes on to occupy a broader convening role within the global AI governance regime complex, as many commentators and governments have been calling for.*

in updating and refining its AI Readiness Assessment, which helps countries understand how prepared they are to apply AI ethically and responsibly. This would ensure the assessment reflects the latest technical developments and emerging risks, especially those associated with frontier AI.

Leveraging the UN Network to Make the AISI Network More Inclusive

The AISI Network could leverage UN platforms to enhance inclusivity, collaborating with agencies like the International Telecommunication Union (ITU) to guide member states in establishing their own AISIs. The ITU, drawing on its expertise in global technology standards, could develop a framework for AISIs that outlines best practices and coordination mechanisms. This partnership would support UN member states in aligning their AI safety initiatives with the global AISI Network, fostering consistent—or at least interoperable—standards and capabilities across nations. AISIs could further contribute to capacity-building efforts by supporting UN agencies such as UNDP, and UNESCO in collecting diverse datasets and empowering member states to use this data to develop model evaluations that reflect specific regional needs and considerations, such as focus on specific areas of risk.

Collaborating with the OECD

The OECD, known for its AI Principles, Policy Observatory, and expert groups, has established itself as a leading authority on technical guidance for AI governance. Its influential role in the multilateral AI governance ecosystem is exemplified by its leadership in supporting and monitoring the G7's Hiroshima Process, the integration of the Global Partnership on AI with OECD AI, and the recently announced enhanced collaboration agreement with the UN.

These developments position the OECD as a strategic partner for coordinated efforts with the AISI Network. AISIs can harness the OECD's analytical expertise, extensive network, and proven ability to translate complex technical insights into actionable policy frameworks, strengthening their capacity to tackle and communicate global AI safety challenges.

Potential collaboration avenues include memoranda of understanding, joint events, and reciprocal advisory roles. The OECD's expert groups (e.g., on data privacy or compute and climate change) and public consultations (e.g., on risk thresholds for advanced AI systems) can offer the AISI Network valuable multistakeholder perspectives. Furthermore, the OECD could consider establishing a dedicated expert group to develop strategies and best practices for AISIs, generating outputs that directly inform the Network's policies and operational frameworks.

Promising areas of collaboration between AISIs and the OECD include:

Coproducing the International Scientific Report on the Safety of Advanced AI

The AISI Network and OECD could collaborate to produce the annual *International Scientific Report on the Safety of Advanced AI*, building on the UK-launched assessment model and strengthening it with the OECD's expertise in rigorous, policy-relevant international reporting.

> *AISIs can harness the OECD's analytical expertise, extensive network, and proven ability to translate complex technical insights into actionable policy frameworks, strengthening their capacity to tackle and communicate global AI safety challenges.*

This partnership would enable the report-writing process to draw from the OECD's experience in publishing yearly, technical, expert-led reports tracking global issues such as its Economic Outlook and the AISI Network's specialized knowledge of AI safety and access to cutting-edge models. Additionally, the OECD's "Key Partner" relationships with major non-member countries like China, India, and Brazil could extend the report's inclusivity and impact, ensuring it captures diverse perspectives on AI safety beyond its 38-member states. This would be particularly valuable for engaging regions where AI governance priorities may differ from Western-centric approaches.

By leveraging the OECD's robust reporting infrastructure and the AISI Network's technical expertise, this collaboration could produce an annual report that complements the broader AI risk assessments expected from the UN's International Scientific Panel and Global Policy Dialogue on AI. The resulting document would provide specialized, frequent insights into advanced AI risks, offering a flexible and timely tool for addressing the field's rapid evolution.

Contributing to the recent OECD and UN Partnership with Specialized Expertise

The AISI Network could play a strategic role in bolstering the recently announced partnership between the OECD and the UN by contributing its specialized expertise in advanced AI safety. This would complement the UN's convening power and inclusive mandate with the OECD's experience in multi-stakeholder policy development and its proven ability to bridge technical and policy communities.

The AISI Network's participation could enhance the partnership's credibility by ensuring that cutting-edge technical insights are effectively integrated into global AI governance efforts. For example, AISIs could contribute nuanced analyses of frontier AI risks, propose actionable safeguards, and provide recommendations on monitoring and evaluation systems.

Such a collaboration could form the foundation of a global AI governance framework that balances technical precision with widespread international adoption. By aligning the strengths of these three entities—the AISI Network's technical rigor, the OECD's policy expertise, and the UN's global reach—this partnership could accelerate the development of governance structures capable of addressing both immediate and long-term challenges posed by advanced AI systems.

Monitoring the G7 Hiroshima Process Code of Conduct:

The AISI Network could assist the OECD in its role of monitoring the G7's Hiroshima Process AI Code of Conduct. **While the Code provides valuable high-level principles, the AISI Network's collaboration could help address its gaps by offering much-needed technical specificity and operational insights.** Key areas of contribution could include developing standardized evaluation frameworks for AI systems, creating robust risk management protocols tailored to frontier AI technologies, and defining clear capability and risk thresholds to guide policy implementation. The Network could also assist in formulating detailed testing and auditing standards, ensuring that the principles outlined in the Code translate into actionable, measurable practices.

By aligning the strengths of these three entities—the AISI Network's technical rigor, the OECD's policy expertise, and the UN's global reach—this partnership could accelerate the development of governance structures capable of addressing both immediate and long-term challenges posed by advanced AI systems.

Incident Monitoring

The OECD's AI Incidents Monitor (AIM) documents AI incidents and hazards, offering insights into risks that could serve as a crucial tool for AI safety. The AISI Network could contribute incident reports to AIM, standardize reporting methods, and verify incidents reported by third parties, adding rigor to global AI risk assessments. Additionally, the AISI Network could lend expertise to further develop the AIM's classification algorithms, ensuring global consistency and interoperability in incident monitoring.

Expanding Collaboration with Other International Coalitions

The AISI Network should also actively engage and collaborate with other multilateral and regional organizations, **especially those that may lack the immediate capacity or intent to establish dedicated AISIs but still have a vested interest in the safe development of AI systems.**

For example, partnerships with the China-BRICS Artificial Intelligence Development and Cooperation Center could offer valuable non-Western perspectives on AI governance, enabling a more inclusive understanding of AI risks and priorities. Such collaboration could also provide a platform for fostering dialogue and cooperation between geopolitical blocs, reducing fragmentation in global AI governance.

Partnerships with the African Union could focus on capacity building, supporting member states in developing the technical and institutional expertise necessary for responsible AI adoption. These partnerships could also prioritize ensuring equitable access to safe and beneficial AI technologies, addressing the digital divide and empowering African nations to shape global AI governance discussions. **The inclusion of the European Union as a member of the AISI Network already offers a precedent for "regional AISIs," or groups of countries collaborating to ensure representation in AI safety discussions without establishing their own domestic institutes.**

Collaborating with multilaterals such as the G7 and G20 could help translate high-level principles into actionable frameworks, particularly in areas like AI risk management and international coordination. Similarly, partnerships with the Council of Europe and ASEAN could enable the development of regionally tailored approaches that respect cultural, legal, and economic diversity while reinforcing global AI safety standards.

By engaging with a wider array of partners across regions, the AISI Network can broaden its influence, integrate diverse viewpoints, and promote a more balanced and inclusive framework for AI safety and governance. These initiatives would build stronger collaboration and mutual trust among stakeholders, driving the safe and ethical advancement of AI technologies on a global level.

Partnerships with the China-BRICS Artificial Intelligence Development and Cooperation Center could offer valuable non-Western perspectives on AI governance, enabling a more inclusive understanding of AI risks and priorities.

6. Conclusion

The formation of the AISI Network is an ambitious step toward building a cohesive, global response to the safety challenges posed by advanced AI. But sustaining the network's effectiveness will require careful planning, adaptability, and a commitment to nurturing trust among its members and partners. As political landscapes shift and AI continues to evolve, the AISI Network should remain a flexible, responsive, and inclusive organization that adapts to new challenges while staying true to its mission.

In this effort, countries worldwide should prioritize support for initiatives like the AISI Network, advocating for equitable contributions and transparent protocols that reinforce trust. The Network must also carefully calibrate its partnerships with AI companies, in order to leverage their invaluable expertise and access to advanced models while safeguarding against potential industry capture. Just as importantly, civil society should find a place within the AISI network, serving as a critical voice for accountability and representing the global public interest.

As the Network prepares to convene in San Francisco, it has a unique opportunity to lay the foundations of a global governance regime that not only adapts to the evolving AI landscape, but also shapes it responsibly, setting the course for a future where AI serves humanity's highest aspirations.

Acknowledgements

A huge thank you to The Future Society colleagues—Nicolas Moës, Yolanda Lannquist, Mai Lynn Miller Nguyen, Toni Lorente, Emily Gillett, and Amin Oueslati—for their invaluable contributions, as well as to our external collaborators: Amanda Leal, Christopher Covino, Jakob Mokander, Kevin Zandermann, Konrad Seifert, and Renan Araujo. We also greatly appreciate the insights gained through informal conversations with stakeholders from the U.S. AISI, the UK AISI, the European Union, and the OECD, which enriched our report.

Participation in this research does not imply endorsement of the report or its findings. The views expressed herein do not necessarily represent the perspectives of these individuals or their respective organizations. Any remaining errors are our own.

Appendix

Source List for Network Diagram of Collaborations Between AISIs and AI Companies

Actors	Type of collaboration
US <> SIngapore	Joint Research Projects
	Personnel Exchanges and Capacity Building
	Harmonizing Regulatory Frameworks
US <> UK	Model Testing and Evaluation
	Personnel Exchanges and Capacity Building
	Joint Research Projects
US <> Anthropic	Model Access and Testing
	Developing Guidelines and Standards
US <> OpenAI	Model Access and Testing
	Developing Guidelines and Standards
UK <> Singapore	Joint Research Projects
	Developing Guidelines and Standards
	Model Testing and Evaluation
UK <> Canada	Joint Research Projects
	Personnel Exchanges and Capacity Building
	Sharing compute resources
US <> Japan	Harmonizing Regulatory Frameworks
US <> Google	Developing Guidelines and Standards
US <> Canada	Joint Research Projects
Japan <> EU	Harmonizing Regulatory Frameworks

Actors	Type of collaboration
EU <> Google	Developing Guidelines and Standards
EU <> OpenAI	Developing Guidelines and Standards
Singapore <> Anthropic	Model Access and Testing
UK <> Anthropic	Model Access and Testing
UK <> Google	Model Access and Testing
UK <> Open AI	Model Access and Testing
UK <> France	Joint Research Projects

THE FUTURE SOCIETY

Contact Us

GENERAL: info@thefuturesociety.org
PRESS: press@thefuturesociety.org

www.thefuturesociety.org