IGAP PRIMER:

AI SAFETY INSTITUTES

FEBRUARY 2025

Background

Recent strides in Artificial Intelligence (***AI***) technologies have resulted in the proliferation of a large number of AI-driven use cases across multiple sectors. This is especially true of highly capable general-purpose AI models which can perform a variety of tasks. This phenomenon has forced a re-evaluation of existing laws, regulations, and standards across the world. As models for AI governance continue to evolve, AI Safety Institutes (***AISIs***) have emerged as a critical mechanism for fostering responsible development, deployment, and testing of AI systems. These institutes serve as hubs for interdisciplinary collaboration, research, and policy guidance, aiming to address potential risks while maximizing the benefits of AI technologies. The information below provides an overview of the purpose, structure, and key functions of AISIs, highlighting their role in fostering safe AI development.

What is AI Safety?

As of January 2025, reasonable international consensus has been established that AI should be designed, developed, deployed, and used, in a manner that is 'safe'. This means it should be human-centric, trustworthy and responsible. However, as noted by the International Scientific Report on the Safety of Advanced AI, our understanding of how general-purpose AI systems work is limited. This is particularly true for frontier AI technologies, including foundation models. Building a shared scientific and evidence-based understanding of significant risks, whether they stem from misuse, unintended consequences, alignment with human intent, or domains like cybersecurity, is crucial to incorporating measures that enhance safety.

'AI safety' is an emerging field of study dedicated to mitigate and prevent the harmful consequences emanating across the lifecycle of AI systems. The field incorporates a combination of government policies, operational practices, principles, and technical mechanisms to achieve this objective. It should be noted that AI safety has overlaps with the components of 'AI Ethics'. Crucial components of AI Ethics include transparency and explainability, fairness, accountability, regulation, protection of human rights, bias mitigation, privacy and data protection.

In May 2024, during the AI Seoul Summit 2024, representatives from Australia, Canada, the European Union, France, Germany, Italy, Japan, South Korea, Singapore, the United Kingdom and USA affirmed the requirement for international collaboration to advance the science of AI safety, and further responsible AI innovation. This followed from the consensus of the previous Bletchley Declaration by a larger group of countries (including India) during the AI Safety Summit in November 2023. These international discussions on AI Safety have compelled some countries to invest in public sector capability for AI testing and safety research. As a consequence of these international commitments, AISIs are emerging globally to help develop AI safety guidance and testing for commercially and publicly available AI systems. However, AISIs being established across the globe are not identical in their design or function. This primer provides an overview of the crucial functions and activities being undertaken at these various institutes.

Functioning of AISIs Globally

A number of countries have already established dedicated and specialized institutions to carry out the role of AISIs. For the development of interoperable standards and evaluation criteria on AI Safety, collaboration and knowledge sharing through international networks will be crucial for the successful functioning of AISIs. Hence, the role played by the International Network of AISIs (***AISI Network***) is important to note before discussing individual institutes. The AISI Network's inaugural convening took place on 20 November, 2024 under the aegis of the United States Department of State and Department of Commerce. This network is intended to harness the capabilities of domestic AISIs and spearhead global efforts to advance the science of AI safety, while also enabling international cooperation on research, best practices, and evaluations.

The AISI Network also echoes the Seoul Statement of Intent toward International Cooperation on AI Safety Science, made earlier in May, 2024. Membership of the AISI Network includes Australia, Canada, the European Union, France, Japan, Kenya, South Korea, Singapore, the United Kingdom, and the United States. However, not all member countries have established domestic AISIs to contribute towards the network.

1. United Kingdom (*UK*)

Date of Establishment: 2 November 2023

One of the foremost AISIs was established by the United Kingdom under its Department of Science, Innovation and Technology. This has been set-up as a research-oriented organization to test advanced AI systems, foster collaboration, and impact global AI development policy. Significant resources have been allocated to the institute, including GBP100 million in initial funding (approximately USD 126 million). The AISI also possesses priority access to AI models from leading companies (including OpenAI, Google DeepMind and Anthropic).

Core Functions:

- Risk Research: The UK's AISI is primarily focused on identifying and mitigating risks associated with AI. It is engaged with a network of researchers to collaborate on key priorities (safeguards, science of evaluations, safety cases and systemic risks) associated with AI. These engagements also incorporate research grants on topics of interest to generate greater scholarship.
- AI Model Evaluation: The AISI's evaluations are focused on identifying the capabilities of advanced AI systems, and risks associated with them. Using in-house capabilities, UK's AISI intends to conduct rigorous safety assessments of advanced AI systems prior to their launch. Crucially, the AISI is engaged in building and running evaluations for:
    - Misuse (scope for dual-use cyber, chemical and biological application of AI)
    - Autonomy (human interaction and autonomous decision-making)
    - Safeguards (robustness of the system's safety and security features against circumvention)
    - Societal impact (impact of the system on democracy, individual welfare and inequality)
- International Collaboration: The AISI coordinates with the wider research community, developers and other governments to impact AI development, and shape global policymaking. Specific focus is placed on countries serious about addressing the risks of AI. The AISI Network is also expected to play a key role in the execution of this function. UK's AISI, in collaboration with governments and private sector AI companies, also intends to produce a set of global standards on AI safety.

UK's AISI has also open-sourced its testing framework, Inspect. This software library enables testers (in start-ups, academia, AI developers and government) to assess the specific capabilities of individual AI models and produce a score based on the results. Inspect has been made freely available to the AI community to adopt and use.

2. United States (*USA*)

Date of Establishment: 2 November 2023

USA's AISI, housed under National Institute of Standards and Technology (***NIST***), United States Department of Commerce, has been established to advance the science, practice, and adoption of AI safety across the spectrum of risks (including those relating to national security, public safety, and individual rights). The AISI under NIST was established with an initial funding of USD 10 million. The priorities for this AISI are based on the White House's Executive Order 14110, in October 2023, which outlined a comprehensive approach towards safe, secure, and trustworthy AI development. It should be noted that USA's new administration, under Trump, rescinded the earlier Executive Order 14110 on AI safety. The implications of this decision on USA's AISI remain unclear, although some experts argue the decision may diminish USA's role in future AI safety initiatives.

Core Functions:

- AI Model Evaluation: The AISI is engaged in conducting safety evaluations for AI models and systems. To enable this, the AISI entered into formal collaboration with Anthropic and OpenAI. The AISI is also placed to lead the Testing Risks of AI for National Security Taskforce, which is expected to undertake testing of advanced AI models across critical national security and public safety domains (including chemical and biological security, critical infrastructure, and nuclear security).
- Formulating Guidelines: This includes the development of guidelines for content authentication and the detection of synthetic content. The AISI is working to develop science-based and empirically backed guidelines or standards for AI safety, in partnership with more than 280 organizations through an AISI consortium.
- Risk Research: Through the Testing Risks of AI for National Security Taskforce, the AISI will undertake coordinated research across critical national security and public safety domains.
- International Collaboration: As a part of its international collaboration initiatives, USA's AISI is working seamlessly in coordination with UK's AISI under to a Memorandum of Understanding signed between the two countries in April 2024. The two institutes recently conducted a joint pre-deployment evaluation of OpenAI's o1 Model.

3. European Union

Date of Establishment: 21 February 2024

While the contours of a dedicated AISI within the EU region are unclear, the European AI Office has been established within the European Commission, as the center of AI expertise, to help implement the EU's AI Act (the first comprehensive legal framework for AI).

This AI Office is also fulfilling some functions akin to AISIs specified above. However, the European AI Office has been provided with a much broader mandate under its guiding legislation. In this respect, it is distinct from other AISIs, due to the possession of certain enforcement powers.

Core Functions:

- Risk Research: The European AI Office has been tasked with developing tools, methodologies and benchmarks for evaluating the capabilities and reach of general-purpose AI models, and classifying models with systemic risks.
- International Collaboration: As part of its mandate, the European AI Office has represented the EU in technical dialogues with the USA's AISI. The AI Office is expected to promote the EU's approach to 'trustworthy AI' with similar institutions globally and foster international collaboration.
- Legal Compliance: The office is also tasked with implementing and monitoring compliance with EU's AI Act through mechanisms including guidance and guidelines, codes of practice, and investigating infringements of the law.

4. Singapore

Date of Designation: 22 May 2024

Unlike other nations discussed here, Singapore has designated an existing institution, the Digital Trust Centre (***DTC***), to function as its AISI. The DTC was initially established with a funding of SGD 50 million (approximately USD 37 million) to undertake research and development efforts on 'trust technologies'. This was set up in collaboration with the Nanyang Technological University.

Core Functions:

- Risk Research: The AISI is also intended to pull together Singapore's research ecosystem and provide science-based inputs to the country's framework for AI governance. Existing priority areas of research include testing and evaluation, safe AI model design, development and deployment, content assurance, and AI policy.
- International Collaboration: Singapore's AISI is expected to collaborate internationally with other AISIs to advance the science of AI safety. The AISI is party to partnership initiatives (on joint testing, information exchange and safety research) with the European Union and the UK.

5. Japan

Date of Establishment: 14 February 2024

Japan's AISI was established under its Information-technology Promotion Agency, in collaboration with various ministries, as an organization that examines and promotes evaluation methods and standards for AI safety.

Core Functions:

- Risk Research: The AISI is expected to conduct research, examine of criteria, and develop standards relating to AI safety evaluation in collaboration with academia. The efforts of the AISI may also inform changes to Japan's AI Guidelines for Business, released in April 2024.
- International Collaboration: The AISI is expected to work with similar organizations in USA, UK and other countries to devise standards and guidance that contribute to improving the safety of AI development. Already, the institute has engaged with the USA's AISI on aligning terminology relating to AI risk management frameworks, among its other initiatives.

6. Canada

Date of Establishment: 12 November 2024

The Canadian AISI is led by Innovation, Science and Economic Development Canada, a federal institution focused on economic development and investment in the country. The AISI was set up with an investment of CAD 50 million (approximately USD 35 million). However, it also leverages capabilities from the National Research Council of Canada, and the broader Canadian research community through the Canadian Institute for Advanced Research.

Core Functions:

- Risk Research: The AISI is centered around advancing the science of AI safety, particularly in collaboration with international partners, to ensure that governments are well-equipped to understand and mitigate the risks associated with advanced AI systems. Additionally, the AISI funds research through the Canadian Institute for Advanced Research on priority issues relating to AI safety, including the short-term and long-term risks of AI systems.
- International Collaboration: The AISI also leverages its expertise to conduct research in collaboration with partners in the international AI Network. The AISI has directly partnered with UK's AISI for developing AI safety standards through a Memorandum of Understanding.

7. South Korea

Date of Establishment: 27 November 2024

The Republic of Korea's Ministry of Science has recently launched the South Korean AISI, within its Electronics and Telecommunications Research Institute, to address AI safety through initiatives focused on risk identification, evaluation, and mitigation. The institute is expected to foster collaboration and information exchange among industry, academia and research organizations. To achieve this goal, a Memorandum of Understanding has been signed with a consortium of 24 organizations including national industry leaders and universities. As this AISI has been recently established, significant details about its intended mandate and initiatives are limited.

Core Functions:

- Risk Research: The AISI is expected to address technological limitations, human misuse, and loss of control over AI systems, among other safety risks.
- International Collaboration: It is also expected to foster collaboration by utilizing the AISI Network.

8. Australia

While Australia is yet to establish a specialized AISI, the country has a National AI Centre (***NAIC***), which was established in 2021, to accelerate the country's AI industry. Among other things, the NAIC has published a Voluntary AI Safety Standard to provide practical guidance to all Australian organizations on responsible use of and innovation with AI technologies. However, stakeholders in Australia have urged government representatives to create a dedicated AISI in the country. Australia was a signatory to the Seoul Statement of Intent toward International Cooperation on AI Safety Science.

It should also be noted that European Union members **France** and **Germany** were signatories to the Seoul Statement of Intent on AI Safety Science. However, information on specific functioning of AISIs established in these countries for international collaboration on AI safety is limited.

What Value an AISI Provide to India?

Recently, India underwent deliberations and preliminary consultations, led by the Ministry of Electronics and Information Technology (***MEITY***), on the requirement for a domestic AISI. In January 2025, MEITY announced the creation of an **IndiaAI Safety Institute** under the 'Safe and Trusted Pillar' of the IndiaAI Mission. This body is expected to work with stakeholders (academia, industry, startups, and government bodies) through a 'hub and spoke model' to ensure safety, security and trust in AI technologies. Similar to other AISIs, the Indian institute will undertake safety research as well. However, this will be contextualized to India's particular social, cultural and linguistic dynamics. The overview of AISIs above indicates that research and testing of AI systems, alongside international collaboration, are central elements to the functioning of these institutes. The AISI may prove to be an effective tool to help India alleviate domestic concerns on AI safety in the following manner:

Addressing Safety: AISIs around the world are working to address the safety risks recognized under the Bletchley Declaration in 2023 through new testing standards and guidance. The establishment of an AISI would provide India with the opportunity to build capabilities for rigorous AI testing and safety assessments, particularly for frontier technologies like foundation AI models.

Nimble Policy Making: Given the rapid pace of AI technology development taking place globally, traditional regulatory approaches may struggle to keep up with emerging risks and challenges. AISIs provide a dynamic and adaptive framework for AI governance, allowing policymakers to make informed, evidence-based decisions in real time.

International Collaboration: Additionally, the risks associated with AI are not confined by national borders. This makes AI safety an inherently global challenge which requires international cooperation. In this environment, operating in silos would be ineffective in addressing AI risks and ensuring safe deployment. As a new kind of government organization, the AISI can leverage its legitimacy and resources to reconcile global concerns with national AI priorities. Countries have recognized AISIs as a critical mechanism for fostering collaboration across regions, organizations, and disciplines. A domestic AISI may be an effective enabler for international collaboration through the AISI Network, and assist India project itself as a leading voice for the Global South on AI safety.

WHAT IS IGAP ?

The Indian Governance And Policy Project (***IGAP***) is an emerging think tank focused on driving growth, innovation, and development in India's digital landscape. Specializing in areas like AI, Data Protection, FinTech, and Sustainability, IGAP promotes evidence-based policymaking through interdisciplinary research. By working closely with industry bodies in the digital sector, IGAP provides valuable insights and supports informed decision-making. Core work streams include policy monitoring, knowledge dissemination, capacity development, dialogue and collaboration.

For more details visit: www.igap.in

IGAP