AISI
Japan AI Safety Institute

Please refer to the original text for accuracy

National Status Report on AI Safety in Japan 2024

February 5, 2025

Japan AI Safety Institute (J-AISI)

Table of Contents

1. Establishment and its efforts to date

1.1. Basic Information

Japan AI Safety Institute (J-AISI) is an organization to study and promote evaluation methods and standards for AI safety in order to realize safe, secure, and reliable AI. Through AI Safety, J-AISI will both address risks and drive innovation in the use of AI.

AI Safety

State that maintained safety and fairness to reduce societal risks* arising from AI use, privacy protection to prevent of inappropriate use of personal data, ensuring security against risks such as external attack caused by vulnerabilities of AI systems, and transparency by ensuring the verifiability of systems and providing appropriate information, based on the human-centric concept.

*Societal risks include physical, psychological and economic risks

1) AISI Objectives (Roles)

- Provide assistance to the government by conducting research on AI safety, examining evaluation methods, and developing standards. J-AISI supports the government by conducting surveys on AI safety, examining evaluation methods, and creating standards.
- As a hub for AI safety in Japan, J-AISI will consolidate the latest information in industry and academia, then promote collaboration among related companies and organizations.
- In addition, an international consensus will be established by collaborating with AI safety-related organizations in other countries.

2) AISI's Unique Situation

AI is at the cutting edge of a technological field that is changing faster than previously thought possible, and its impact is not limited to technology, but spans society, employment, culture, and much more.

Each country's AISI situation is different, and the scope of AI safety covers a wide range of issues, from the security of the AI system itself, to the governance of the entire ecosystem, laws and other arrangements, to intellectual property and personal data considerations. In the US, the scope has been further expanded to include national security.

The speed of technological development is so fast that products are replaced on a

monthly or quarterly basis, and it is a cross-disciplinary and global effort, with reports published monthly by various countries and international organizations.

In this situation, each initiative in Japan and around the world is taking longer than ever before, and we are truly in a situation where we are thinking as we run, using agile thinking. We are being asked to promote a completely new technology policy, rather than the traditional approach of responding to stable technologies that change on a yearly basis. We are being asked to respond to this not only in terms of the sheer number of technologies, but also by checking the connections between technologies to ensure their quality.

There has been no precedent for cross-national or global coordination efforts on AI safety, and countries are still searching for ways to create a structure that integrates policy and technology. In addition, while each country invests a reasonable amount of money, the way in which each country creates its structure differs, and these differences make coordination even more difficult, while AISI can directly promote policy on AI safety as a country by operating as part of a government agency.

Driving structure of policy and technology in each country

U.S.: State and Commerce Departments and the National Institute of Standards and Technology (NIST) teamed up to promote the initiative.

U.K.: Creation of AISI within the Ministry to unify policy and technology decision making.

European Commission: The Commission is in charge of the law and also promotes the utilization in a unified team within the Organization.

Japan: The Cabinet Office and relevant ministries collaborate to provide government policies and directives to AISI for ensuring AI safety. Based on these directives, AISI conducts investigations related to safety evaluations, examines standards, and considers implementation methods.

1.2. Key events and actions

1) AISI

In February 2024, ten ministries and five government-affiliated organizations jointly established the Japan AI Safety Institute (J-AISI). J-AISI is an organization that supports the government's efforts related to AI safety. Given its extensive information-gathering capabilities, it is expected to serve as a hub for information on AI safety within Japan. Furthermore, as international initiatives on AI safety progress at a very rapid pace, AISI

is also expected to participate in and contribute to such international efforts.

As part of its efforts to provide information and ensure transparency, AISI established a website upon its establishment and completely renewed it in September 2024, reporting on the status of AISI activities and making its deliverables publicly available. It also provides a link to the Cabinet Office on the policy side, a link to Artificial Intelligence R&D Network (AI Japan R&D Network) on the R&D side, and a link to Information-technology Promotion Agency (IPA) on AI utilization, realizing a hub function for information on AI safety.

AISI website

https://aisi.go.jp/

2) AI Guidelines for Business

In April 2024, the Ministry of Internal Affairs and Communications (MIC) and the Ministry of Economy, Trade and Industry (METI) integrated and updated the existing AI guidelines and published them as the AI Business Operator Guidelines. The guidelines have served as a foundation for the development of AI safety, as the financial industry has developed its own AI guidelines and created the Guide to Evaluation Perspectives on AI Safety, which will be discussed later. In addition to the Japanese version, an English version is also available for international collaboration.

Currently, AISI is also participating as a joint secretariat and is working on a continuous update following the 1.01 version published in November 2024.

https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20241226_1.pdf

https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20241226_2.pdf

3) AI safety summit

A summit on AI safety was held in Seoul in May 2024 to deepen the discussion on AI safety, drive innovation in AI development and equitable sharing of AI benefits, and the Seoul Declaration for Safe, Innovative and Inclusive AI was adopted as a summit-level outcome document. The Seoul Declaration for Safe, Innovative and Inclusive AI was adopted as the summit-level outcome document.

AISIs from Japan and other countries participated in the AI Global Forum, which was held in conjunction with the AI Global Forum, and held bilateral meetings to exchange views on the future of international cooperation.

https://www.mofa.go.jp/mofaj/files/100672518.pdf (Japanese Only)

4) Japan-U.S. Crosswalk

In April and September 2024, AISI published the results of a crosswalk (cross-reference of terminology
and concepts) between the Japan AI Guidelines for Business and the US NIST AI Risk Management Framework. This has been conducted since January 2024, prior to the establishment of AISI.

The U.S. NIST AI Risk Management Framework has already been crosswalked with guidelines and international standards of other countries, and cross-referencing with guidelines of other countries is possible via the U.S. NIST AI Risk Management Framework. The crosswalk between Japan and the U.S. has enabled us to confirm the interoperability between Japan and the U.S. on AI risk management by cross-referencing concepts ahead of other countries. Based on this, we have taken an approach to develop the Guide to Evaluation Perspectives on AI Safety and the Red Teaming Methodology Guide on AI Safety.

Crosswalk 1
https://aisi.go.jp/effort/effort_information/240430/
Crosswalk 2
https://aisi.go.jp/effort/effort_information/240918_1/

5) AI Institutional Research Group

This is an expert panel established by the government in August 2024 to study the development of laws, regulations, and institutions for AI. The main topics of discussion include: response to technological progress including generative AI, harmonization with international guidelines, flexible institutional design, etc. The "Draft Interim Summary" was submitted in December 2024, and opinions were solicited. The Executive Director of J-AISI is acting as the deputy chair of the AI System Research Group and is leading the study.

https://www8.cao.go.jp/cstp/ai/ai_kenkyu/ai_kenkyu.html (Japanese Only)

6) Guide to Evaluation Perspectives on AI Safety

In September 2024, based on the AI Guidelines for Business, a guide to evaluation perspectives on AI safety was developed, presenting basic concepts that those involved in the development and provision of AI systems can refer to when conducting AI safety evaluations. In addition to the Japanese version, an English version is also provided from the viewpoint of international collaboration.

Guide to Evaluation Perspectives on AI Safety

https://aisi.go.jp/effort/effort_framework/guide_to_evaluation_perspective_on_ai_safety/

7) Red Teaming Methodology Guide on AI Safety

In September 2024, we developed a guide to the red teaming methodology, an AI safety assessment method, which provides basic considerations for those involved in the development and provision of AI systems to assess the measures taken to address the risks posed to the target AI system from an attacker's perspective. In addition to the Japanese version, an English version is also provided from the perspective of international collaboration.

Red Teaming Methodology Guide on AI Safety

https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/

8) United Nation Global Compact Leaders Summit

In September 2024, the Leaders Summit, an international conference held on the occasion of the UN Future Summit, is an important forum to connect with businesses on a principles-based approach while addressing risks to promote sustainable development and prosperity Expectations for AISI to promote the responsible development and use of AI have increased In response to the growing expectations for AISI to promote the responsible development and use of AI, the executive Director of J-AISI held discussions with a diverse range of stakeholders, including national government AI safety officials and businesses, during the "Rights & Bytes" session to demonstrate actionable steps in human rights and AI and to encourage participants to take action.

https://events.unglobalcompact.org/leaderssummit24

9) CEATEC

In October 2024, the executive director of J-AISI gave a lecture titled "The Future of AI Society" at CEATEC (Combined Exhibition of Advanced TEChnologies) held at Tokyo, as well as an exhibition and lecture at the special "AI for ALL" booth for the 25th anniversary of CEATEC. AISI also exhibited and gave a lecture at a booth specially set up for the 25th anniversary of CEATEC "AI for ALL". In addition, the executive director of J-AISI gave a keynote speech at a networking event on the theme of AI with domestic AI promotion organizations, and exchanged opinions with related organizations.

10) Singapore AI Safety Red Teaming Challenge

In November 2024, Japan participated as a major participant in the International Red Teaming Challenge in Singapore. This event was related to AI safety with a focus on multicultural and multilingual AI. From Japan, the National Institute of Informatics (NII), the University of Tokyo, the Kyoto University, and the Japan Deep Learning Association (JDLA) also participated in the event and conducted an evaluation of large-scale language models.

https://www.imda.gov.sg/activities/activities-catalogue/singapore-ai-safety-red-teaming-challenge

https://www.tc.u-tokyo.ac.jp/blog/wp-content/uploads/2024/10/Singapore-Report-English.pdf

11) Launch of the International Network of AI Safety Institute (AISI)

This is an international cooperation framework involving 9 countries and 1 area—the United Kingdom, the European Commission, Australia, Canada, Kenya, Singapore, the Republic of Korea, Japan, France, and the United States—to discuss the technical aspects of AI safety. The first meeting was held in San Francisco in November 2024, where the mission statement was adopted. Additionally, discussions were conducted on "Track 1: Reducing Risks of Synthetic Content," "Track 2: Joint Testing of Foundational Models," and "Track 3: Risk Assessment of Advanced AI Systems." In the session considering evaluation methods in multiple languages, Japan co-chaired with Singapore.

Mission Statement

https://aisi.go.jp/effort/effort_international/241120/

1.3. Other AISI-related activities and accomplishments

1) Japanese translation of the U.S. NIST AI Risk Management Framework and Playbook

In July 2024,In order to deepen understanding of AI safety in Japan, we have created a Japanese translation of the AI Risk Management Framework developed by the US NIST, which is used as an important document for risk management in international settings, not just in the US, and we have gone through a process of checking with the US government, including the Department of Commerce and the Department of State, to create a fixed interpretation between Japan and the US.

https://aisi.go.jp/effort/effort_information/240704/

2) Security Research Report

With AI being used in business as a new technology, AISI and IPA conducted a survey to understand the actual status of security threats and risks in Japan and overseas, and published the results of "Survey on Perceived Security Threats and Risks of AI in the United States" in May 2024, "Survey on Security Threats and Risks When Using AI Report" in May 2024, and "Election Risks Caused by AI and AI Governance U.S. Survey" in December 2024.

https://aisi.go.jp/effort/effort_technology/technology/

3) Survey on AI Policy Trends

In May 2024, we published the "Survey of Recent Trends in AI Policy in the United States" to understand the latest trends in AI policy in the United States. Focusing particularly on the efforts of the U.S. National Institute of Standards and Technology (NIST), the survey included a summary of NIST's major AI-related efforts to date, an overview of AI-related research at NIST, requirements for NIST under Presidential Executive Order (No. 14110), and a summary of the U.S. AI Safety Institute (USAISI) and the U.S. AI Safety Institute Consortium (AISIC), both of which are under NIST. The U.S. AI Safety Institute (USAISI) and the AI Safety Institute Consortium (AISIC), both of which are affiliated with NIST, were outlined.

https://aisi.go.jp/effort/effort_technology/technology/

4) Added supplemental information on generative AI and other topics to the "DX Promotion Skill Standards" for personnel promoting DX.

In July 2024, the IPA revised the Digital Skill Standards, including addressing the risks of AI. The characteristics of generated AI, examples of its use, and how to respond to new technologies are supplemented, and learning items on the Common Skill List are added or changed. The Digital Skill Standards consist of two parts: the DX Literacy Standards, which are required of all business people, and the DX Promotion Skill Standards, which define the skills and roles required for DX promotion.

5) Activity Map and Terminology Challenge

The Activity Map visualizes the overall picture of overlooked areas and correlations between activities as the activities related to AI safety continue to rapidly change and evolve. Japan is developing a comprehensive activity map and related terminology based on benchmarking of key literature. This Japan-led effort is expected to further strengthen the foundation for international collaboration in AI safety and contribute to

the realization of a sustainable and trustworthy AI society. (To be published in February 2025)

6) Consideration of a data quality guide for AI systems

Data quality management is essential for the appropriate and safe use of AI. Therefore, we have studied a useful, practical, and pragmatic guide for companies, while utilizing existing international standards and guidelines on data quality. Data quality is one of the evaluation perspectives in the Guide to Evaluation Perspectives on AI Safety. In this, the handling of AI-specific data, such as synthetic content, is also examined. (Scheduled to be published in February 2025)

7) Consideration of Joint Certification for AI Certification

ISO/IEC JTC 1/SC 42 - Artificial intelligence has decided to develop a high-level framework standard for conformity assessment of products jointly with the senior committee developing conformity assessment standards in ISO (CASCO). In the course of this development, a new certification framework, Joint Certification (a new method of combining organizational certification and AI product/service certification), will be considered in parallel. Therefore, the policy for dealing with this issue was discussed.

8) Consideration on the appropriate procurement and use of AI by Japan government

AISI participated in discussions on appropriate procurements and utilizations of AI by Japanese government, led by the Digital Agency and other governmental bodies, and contributed to the discussions by making proposals for addressing the risks associated with generative AI.

Japanese government will formulate guidelines for the procurement and utilization of AI in the governmental organizations approximately by the spring of 2025.

9) International Activities

AISI has exchanged views with AI stakeholders in various countries in person and online, and has also given lectures and conducted training lecturers (total 14 times).

Main meeting partners

- Discussions with Australia, Canada, European Commission, France, Japan, Kenya, Malaysia, Malta, Republic of Korea, Singapore, the United Kingdom, United States
- Opinion exchange with overseas AI-related organizations

- JICA training (Armenia, Yemen, Uganda, Grenada, Zambia, Tonga, Vanuatu, and Jordan), and joint training by the ASEAN-Japan Cybersecurity Capacity Building Centre (AJCCBC) and JICA (AJCCBC seminars are for government officials and critical infrastructure officials from ASEAN member countries, with over 100 participants)

10) Domestic Dissemination Activities

As part of domestic AI safety awareness activities, we have given lectures and participated in panel discussions and at the Japan Business Federation (Keidanren), AI Governance Association, Tokyo Metropolitan SME Intellectual Property Symposium, LLM Symposium, and AI Quality Management Symposium (total 60 times).

1.4. AISI Structure

1) Framework for Collaboration on AI Safety

In cooperation with the Cabinet Office, a framework for collaboration among 10 relevant ministries, agencies, and 5 related organizations has been established; a liaison conference of AISI-related ministries, agencies, and others and a steering committee have been established to discuss policies.

[Relevant government ministries and agencies]

Cabinet Office (Secretariat of Science, Technology and Innovation Policy),
National Security Secretariat, National Center of Incident Readiness and Strategy for Cybersecurity, National Police Agency, Digital Agency, Ministry of Internal Affairs and Communications,
Ministry of Foreign Affairs of Japan, Ministry of Education, Culture, Sports, Science and Technology, Ministry of Economy, Trade and Industry, Ministry of Defense

[Related organizations]

National Institute of Information and Communications Technology (NICT), RIKEN, National Institute of Informatics (NII), National Institute of Advanced Industrial Science and Technology (AIST), Information-technology Promotion Agency (IPA), Japan

2) Japan AI Safety Institute Partnership

In August 2024, J-AISI concluded a partnership agreement with five related organizations (National Institute of Information and Communications Technology, RIKEN, National Institute of Informatics, National Institute of Advanced Industrial

Science and Technology, and Information-technology Promotion Agency, Japan) to establish domestic cooperation on AI safety. In order to effectively drive J-AISI activities, J-AISI related organizations that J-AISI have knowledge about AI safety and jointly conducts research and surveys, and disseminates information domestically and internationally.
https://aisi.go.jp/effort/effort_information/240806/

3) AISI Secretariat

The Secretariat is located within the IPA and consists of five teams.

- Strategy and Planning Team
  - Coordination with government agencies
  - Coordination with private sector activities
  - International Coordination
  - Project Management
  - General Planning
- Framework Team
  - Survey
  - Guideline development
  - Framework International Coordination
- Technical Team
  - Survey
  - Activity map development
  - Technical Verification
  - International Technical Coordination
- Security Team
  - Survey
  - International Security Technical Coordination
- standard team
  - Survey of trends in international standards (including activities in ISO/IEC JTC1/SC42 (Artificial intelligence))
  - Standard International Coordination

As of January 2025, there are about 30 people, including full-time and part-time. As the scope required for AI safety expands, a small group of experts has been working on the cutting edge of AI safety activities.

1.5. Challenges due to AISI's unique situation

AISI is organized as a technical expert team on AI safety, but from the user perspective, it is easier to understand and there is a demand for an integrated approach to include the use of AI, rather than just a deep dive into AI safety. There is also a demand to address data quality described in Appendix 3. For AI Developers of the AI guidelines for business, "Proper data training", which includes "Verify that the data does not contain personal data, confidential information, rights including copyrights or legally protected interests". Taking into account the changing international landscape such as the AI Action Summit which is considered to be a step forward from the conceptual level discussions at previous AI Summits, it is necessary to consider with stakeholders how to address areas that cannot be solved by technology alone as an AI safety.

AISI is working proactively to secure human resources in the field of AI in cooperation with domestic partners, as it is required to respond to a wide range of stakeholders, including counterparts in government ministries, engineers, and overseas. However, the current shortage of human resources in AI field both domestically and internationally is challenging, due in part to restrictions on the treatment of civil servants.

In addition, the governance structure in Japan is multi-layered and difficult to be flexible and agile, as there are many agencies involved and the virtual organization must operate under different information management policies and environments.

In addition, the OECD has identified 10 potential future AI risks in "Assessing potential future Artificial Intelligence risks, benefits and policy imperatives" published in November 2024. In addition to technological risks, it points to the risk that "Governance mechanisms and institutions unable to keep up with rapid AI evolutions."

2. For the future

2.1. New factors to consider

In addition to the issues in 1.5, factors to consider for the future include the following

1) Evaluation methods and criteria for AI safety

1-1) Results of AI Institutional Research Group

The results of the study group's discussions are expected to various organizations

influence future AI safety structures and promotion methods of. For example, clarifying rules and standards for AI use is expected to improve AI safety and ethics and increase transparency. Although the application of regulations may be burdensome, the increased reliability is expected to promote technological development and investment, thereby enhancing competitiveness in certain sectors. In addition, enhanced risk countermeasures, such as protection of personal information and control of bias, are expected to create an environment in which citizens can use AI with peace of mind, and AI literacy is expected to improve. Furthermore, consistency with international standards is expected to facilitate international cooperation and market entry, and promote society-wide efforts to adapt to the AI era.

1-2) Trends in Certification Processes

ISO/IEC 42001 is an international standard that specifies requirements for the establishment, implementation, maintenance, and continuous improvement of Artificial Intelligence Management Systems (AIMS) within organizations. It is designed for businesses that provide or use AI-based products or services, and ensures the responsible development and It ensures responsible development and use of AI systems.

Currently, ISO/IEC 42006, which specifies requirements for certification bodies, is being developed in preparation for the launch of a certification system based on ISO/IEC 42001. These international developments are important for companies and organizations to remain competitive in the global marketplace and should be addressed in Japan. It is imperative to keep a close eye on the certification trends and consider appropriate responses.

2) Technical Progress

New AI safety issues may emerge as a result of the expansion of new technologies and applications of AI. New technologies related to AI may also change the scope, perspectives, and evaluation methods of AI safety. The importance of AISI activities will increase in various areas, including the need to update AISI-issued guides in light of these technological developments.

3) Collaboration with AI safety agencies in other countries

3-1) Effort by International Network of AISI

The future direction of the International Network of AISI will be significantly

influenced by the discussions at the AI Action Summit to be held in Paris in February 2025. Whether to continue developing the current themes or to establish new ones may determine the path forward.

3-2) Efforts by International Organizations, etc.

As discussions on AI are underway at the G7 and at international organizations such as the OECD and the UN, and the content, promotion structure and specific methods of these discussions are expected to have an impact both domestically and internationally. These discussions revolve around themes such as the ethical use of AI, the state of regulations, and the strengthening of international cooperation, and may have a significant impact on the policies and systems of each country, including Japan. In particular, the formation of international agreements and standards will be an important factor in determining the direction of domestic AI strategies and regulations.

4) Collaboration with private sector

AISI In the UK and the US is working with the private sector to evaluate AI models and develop evaluation tools. This effort is not only to pursue the accuracy of the technology, but also to ensure the reliability and safety of AI. These assessment tools are also designed with a strategic perspective to address long-term impacts and risks, taking into account their utility for industry and society as a whole.

2.2. Future Direction of Efforts

Based on the main challenges and new elements to be considered for AISI that are currently visible, we will enhance and upgrade the activities that contribute to AISI's founding purpose through the following initiatives to achieve AI safety.

1) Research on AI safety, study of evaluation methods and enhancement and upgrading of guidelines

- In the international context where new technologies and threats to AI are emerging one after another, we will investigate and revise the Guide to Evaluation Perspectives and the Guide to Red Teaming Methodology published by the AISI to make them more relevant to the actual situation.
- A feasibility study on technologies to automate/laborsaving will be conducted with the aim of widely disseminating AI safety assessment activities to the general public.
- Focus on common technical areas such as data quality and Joint Certification.

2) Measures to deal with technological progress

- Expand the scope to include highly versatile multimodal foundation models and conduct research on cutting-edge use cases/technological trends that close to practical application.
- Revise the guideline by detailing the evaluation methods/items by practicing the "Evaluation Perspective Guide" and "Red Teaming Methodology Guide" published by J-AISI.
- As public assistance, we will work to improve AI safety in society as a whole through the provision of educational materials.

3) Collaboration with AI safety organizations in other countries

- Promote international collaboration and coordination through AISI-related top-level cooperation and exchange of views with other countries. For example, based on the joint statement on the key points of risk assessment for advanced AI systems, J-AISI will continue to work with AISI in each country to develop a breakdown into specific actions and evaluation methods.

4) AISI's measures for unique situations and cooperation with the private sector

- AISI will drive advanced initiatives and build a sustainable structure by acquiring key players in policy, technology, and international coordination. One of these is the exchange of personnel between the public and private sectors. Secondees from the private sector are active within AISI, and we will continue to recruit personnel to strengthen this and the structure.
- In order to take prompt and appropriate action, AISI will establish a new information-sharing infrastructure that will handle policy and technology across the board, taking advantage of the diversity of AISI-related organizations. Actions will be taken such as putting the shared information into a pull form for each stakeholder to actively utilize it.
- In the early stage of 2025, working groups will be established to promote the development and demonstration of technologies related to AI safety, and to promote studies on specific issues in each industry sector, mainly by the private sector. The working groups will promote applications in priority fields such as industry, healthcare, and disaster prevention, etc. It will formulate and develop guidelines, evaluation tools, evaluation data sets, etc. to ensure AI safety, and contribute to the expansion of the information hub function for AI safety in Japan.