AISI
Japan AI Safety Institute

Please refer to the original text for accuracy

# National Status Report on AI Safety in Japan 2025

March 31, 2026

Japan AI Safety Institute (J-AISI)

# Table of Contents

# 1 Establishment of the Japan AI Safety Institute (J-AISI) and Past Initiatives

## 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

A state which based on a human-centric approach, safety and fairness are maintained to reduce societal risks* arising from the use of AI; privacy is protected to prevent the inappropriate use of personal data; security is ensured to address risks such as vulnerabilities in AI systems and external attacks; and transparency is maintained by ensuring system verifiability and providing appropriate information.

*Societal risks include physical, psychological, and economic risks.

### 1) AISI Objectives (Roles)

- Assist the government by conducting research on AI safety, examining evaluation methods, and developing 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.

The institutional context surrounding AI safety differs from country to country, and the scope of AI safety covers a wide range of issues, from the security of the AI systems themselves to the governance of the entire ecosystem, legal and regulatory frameworks, and considerations related to intellectual property and personal data. 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, adopting an agile approach. 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.

Policy and technical governance structures in selected jurisdictions

United States: The Department of State, the Department of Commerce, and the National Institute of Standards and Technology (NIST) jointly promote these initiatives.

United Kingdom: 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

In light of changes in the international and technological environment, J-AISI developed its activities in FY2025 with strategic emphasis on (1) systematizing evaluation methods and guidelines, (2) responding to systems, standards, and certifications, (3) deepening international cooperation, and (4) responding to industrial and social development.

### 1) Systematization of evaluation methods and guides

J-AISI does not limit AI safety to an abstract philosophy but emphasizes the systematic development of evaluation methods and guidelines that can be applied to practice.

#### (1) Release of the Guide to Evaluation Perspectives on AI Safety (Version 1.10) (Revision and Expansion of Multimodal Perspective)

The guide, which summarizes the perspectives of evaluating AI safety (evaluation perspectives related to AI safety), has been revised in light of changes in technological trends and released as version 1.10 in March 2025. In the revision, the survey results and examples of evaluation items related to multimodal were added. By providing it as a common language and check perspective that can be used for evaluation design and internal and external explanations, the aim is to make it easier for businesses to incorporate safety evaluation into practice.

https://aisi.go.jp/output/output_framework/guide_to_evaluation_perspective_on_ai_safety/

#### (2) Publication of the Guide to Red Teaming Methodology on AI Safety (Version 1.10) (Examples, Attachments/Revisions with Attachments)

The Guide to Red Teaming Methodology on AI Safety, which was published in 2024, has been revised to support the practice of business operators and released as version 1.10 in March 2025. In the revision, based on the results of the implementation of the prototype AI system, detailed procedures and reports were added as attachments. By clarifying how the methodology can be operationalized, rather than merely presenting a conceptual overview, the guide provides reproducible operational procedures that can be applied in practice.

https://aisi.go.jp/output/output_framework/guide_to_red_teaming_methodology_on_ai_safety/

(3) Development of AI safety evaluation tool (AI safety evaluation environment) and release of OSS

For businesses involved in the development and provision of AI systems to practice AI safety evaluation based on the “Guide to Evaluation Perspectives on AI Safety” and the “Guide to Red Teaming Methods for AI Safety”, it is necessary to incorporate the contents described in the guide into specific evaluation items and datasets before conducting evaluation tests on AI systems. The need for specialized knowledge and a process of trial and error to achieve this was an issue for promoting the spread of AI safety evaluation activities. For this reason, we developed an AI safety assessment environment as a reference implementation of an evaluation tool and dataset that embodies the guide-based AI safety assessment, and released it to the public as open source software (OSS), including the source code and dataset of the tool (September 2025).

In addition to being available as a free evaluation tool and dataset, the AI safety assessment environment is also intended to be used to support various businesses in developing customized evaluation tools and datasets according to their own environments and conditions by referring to source code and datasets.

Furthermore, in discussions and technical exchanges with AI safety institutes and government agencies in other countries, Japan will be able to present its position and approach on AI safety more concretely by presenting the tool together with the guides, and the evaluation environment is expected to contribute to future international cooperation.

https://aisi.go.jp/output/output_information/250912/
https://github.com/Japan-AISI/aisev

(4) Development of an evaluation tool in public-private collaboration by the AI Safety Evaluation Environment Task Force

After releasing the AI Safety Evaluation Environment in September 2025, J-AISI established the “AI Safety Evaluation Environment Task

Force“ in October 2025, with participation from approximately 10 private companies engaged in AI-related businesses. The Task Force was established to gather feedback on the AI Safety Evaluation Environment and discuss directions for future functional enhancements.

At the Task Force, participating companies shared perspectives on how their AI business challenges aligned with, or differed from, the current AI Safety Evaluation Environment. The discussions helped identify issues related to AI safety and the evaluation environment, as well as potential directions for future enhancements, providing a basis for developing specific enhancement proposals from FY2026 onward.

In addition, to widely share the FY2025 discussions and outcomes with stakeholders beyond Task Force participants, J-AISI held the “AI Safety Evaluation Environment Task Force General Meeting“ in February 2026 as an event open to the public. Approximately 100 people participated both in person and online. Through presentations by J-AISI and Task Force participants, the event helped raise awareness of and interest in the evaluation environment and the Task Force.

https://aisi.go.jp/activity/activity_information/260306/ (Japanese Only)

(5) AI Incident Response Approach Book Released

An approach book showing “AI-IRS (AI Incident Response System)” as a new framework for responding to incidents caused by risks unique to AI systems has been released. The book assumes that AI incidents are possible and clarifies the importance of heuristic controls to minimize damage. The AI-IRS presents a concept that takes into account the entire AI supply chain, offering an approach to leveraging AI as a reliable social infrastructure. (Released in January 2026)

https://aisi.go.jp/activity/activity_security/260109/

(6) Known Attacks and Their Impacts on AI Systems

In order to provide a bird's-eye view of security attacks unique to AI systems, we have summarized and published attacks on AI and AI systems published in academic papers and other publications as “Known Attacks and Impacts on AI Systems“. (March 2025)

https://aisi.go.jp/output/output_security/known_attacks_and_impacts/

In addition to the use of this document for red teaming, which complements the typical attack methods on AI systems described in the "Red Teaming Method Guide on AI Safety" that has already been published, this document can also be used for studies and research and development on AI security. It was also adopted as reference material at the Ministry of Internal Affairs and Communications' "AI Security Subcommittee" (1st meeting).

In addition, more detailed content was published in the academic paper "Securing AI Systems: A Guide to Known Attacks and Impacts (arXiv:2506.23296)". (June 2025)
https://arxiv.org/abs/2506.23296

(7) Published a data quality management guidebook for AI systems

AI will provide convenient and efficient services and accelerate social change. In order to promote the use of AI, it is essential to ensure trust so that AI can be used with confidence. Data is the foundation of AI, and correct data can be derived by learning and processing the correct data. However, if the data is incorrect, it will be difficult to derive the correct answer, undermining the reliability of the entire process. Data quality is the cornerstone of AI excellence and contributes to the realization of reliable AI.

Based on this background, this guide summarizes what should be done to ensure the data quality necessary to continuously maximize the value of data and AI in order to properly realize an AI society and lead to a data-driven society. (Version 1.01 released in December 2025)
https://aisi.go.jp/output/output_framework/data_quality_management_guidebook/

(8) Report on the Specific Influence on AI Safety

By investigating the specific threat situation that AI systems pose to society and publishing a report summarizing the results, "what and how it can become a problem" was presented as a concrete picture. This survey expands the scope of the survey from the direct impact of AI systems on end users to the impact of AI on society, which has been investigated in the previous guide (Evaluation Perspective Guide on AI

Safety). It is positioned as a basic material for discussing the evaluation design of AI systems on an impact-based basis. (Announced in October 2025)
https://aisi.go.jp/output/output_information/251031/

(9) Publication of the AI Safety Approach Book (document for the spread of AI safety)

The “Evaluation Perspective Guide on AI Safety” and “Red Teaming Methodology Guide for AI Safety” (hereinafter referred to as the “Guides”) published by the Japan AI Safety Institute (J-AISI). We hope that more people will use it to realize AI safety. In order to help those who are referring to both guides for the first time to understand the overview, we have created an approach book (leaflet and PowerPoint) that summarizes the main points in a concise manner. (Released in March 2025)
https://aisi.go.jp/output/output_information/250326/

2) Response to systems, international standards, and conformity assessments

(1) Release of AI Guidelines for Business

In April 2025, the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry announced version 1.1 of the AI Guidelines for Business, which are unified guidelines for AI governance in Japan. Updates were made to organize new risks by AI, reflect the latest technology trends, and reflect domestic and international trends related to AI governance.

The AI Guidelines for Business are a unified guideline for AI governance in Japan for businesses from various positions, including developers, providers, and users, so that those who use AI can correctly recognize the risks of AI and voluntarily implement necessary measures.

J-AISI operates the AI Guidelines for Business Study Group as a joint secretariat with the Ministry of Economy, Trade and Industry and is working on continuous updates. (Version 1.2 released in March 2026)
https://www.ipa.go.jp/disc/committee/expert-group-on-aigfb.html

(2) ISO/IEC JTC 1/SC 42 (AI/Artificial Intelligence) Activities Related to International Standardization

・International

AISI participated in the Spring General Meeting (India) and the Autumn General Meeting (Australia). The next general meeting will be held in Singapore in the spring.

As a major event, a joint working group (JWG7) was established to standardize AI in the financial field. The development of ISO/IEC 42007, which stipulates a high-level framework for the conformity assessment of AI systems, is being carried out in collaboration with ISO CASCO (JWG6). In addition, standards are being developed for functional safety and AI testing technology, including AI and Red Teaming. Japan also led discussions with the SC 42 secretariat on overall operations, and exchanged views with the JTC 1 secretariat on clarifying criteria for establishing joint working groups, given the increasing number of JWGs.

・Domestic

Members of J-AISI's Standards Team served as secretaries for SC 42, WG 1, and JWG 6, supported the operation of the SC 42 National Committee and its subcommittees, and coordinated the review of voting items.

・AIST Project on HMT and AIMS

The AIST project continued its work on HMT (Human-Machine Teaming) and AIMS (AI Management System). As part of its research related to AIMS, the project team exchanged views with stakeholders in Canada and discussed a financial-sector pilot assessment project conducted by the Standards Council of Canada (SCC) and EY.

・Others

Members of J-AISI's Standards Team served as chairs of the drafting committees for JIS Q 42005 and JIS Q 42006. Following the work on JIS Q 42001 in the previous year, they convened six drafting committee meetings for each standard and prepared the drafts with the cooperation of committee members and support from the Japanese Standards Association.

A planning session was held at the National Conference of the Japan Society for Artificial Intelligence (Osaka), where J-AISI was introduced and AI conformity assessment was presented.

(3) Standard BRIDGE-related Activities

"Consideration of Joint Certification in the AI field" was promoted.

Research and studies are conducted on AI conformity assessment. With the goal of formulating a winning scenario for Japan, a committee of experts composed of experts from certification organizations and standards development organizations was formed jointly with the Conformity Assessment SWG, which was established as one of the J-AISI Project Demonstration Working Groups, to organize the existing conformity assessment system and examine the ideal way of conformity assessment in the AI field on the theme of HMT.

As part of the study, J-AISI exchanged views in Luxembourg, the United Kingdom, and Switzerland with the convener of CEN/CENELEC JTC 21/WG 2, UK representatives to ISO/IEC JTC 1/SC 42, and the convener of ISO/IEC JTC 1/SC 42/JWG 6.

In the United Kingdom, Germany, and France, J-AISI exchanged views with representatives from Microsoft, the Alan Turing Institute (ATI), DAkkS (the German accreditation body), GovTech (a non-profit organization that connects government and industry to support the practical implementation of policy in society), and AFNOR (the French standardization body).

The discussions covered topics such as Joint Certification, ISO/IEC 42007, Germany's activities in ISO/IEC JTC 1/SC 42, ISO/IEC 42001 certification in Germany, and the activities of the AI Standards Hub, including its March event. J-AISI also exchanged views with ISO/IEC JTC 1/SC 42 stakeholders in France, and discussed the development status of harmonized standards under the EU AI Act with the Chair and Vice-Chair of CEN/CENELEC JTC 21.

We conducted joint research with Kyoto University and examined LE4SDS. We examined laws and standards (based on the European AI Act and the Harmonized Standards, New Legislative Framework: NLF, etc.), and examined the conformity assessment as evidence that the law is functioning. The study examined the state of law for new technologies

such as AI (from Legal Tech to Legal Engineering) and the state of law enforcement and technology.

3) Deepening international cooperation

(1) Hiroshima Global Forum for Trustworthy AI

The Japan AI Safety Institute (J-AISI, Director: Akiko Murakami) jointly held the “Hiroshima Global Forum for Trustworthy AI” in January 2026 in collaboration with the Artificial Intelligence Policy Promotion Office of the Science, Technology and Innovation Promotion Secretariat, Cabinet Office. This forum was planned as a forum to discuss the latest trends in Japan and abroad to improve the safety, security, and reliability of AI, as well as the efforts of AI safety and security agencies in each country.

https://aisi.go.jp/activity/activity_international/260130/

(2) J-AISI International Network of AI Safety Institutes (Sharing of International Joint Test Exercises and Evaluation Methodologies)

As an international cooperation framework to examine the technical aspects of AI safety, the first meeting of the International Network of AI Safety Institutes was held in November 2024. Since the Paris meeting in February 2025 (the second meeting), the evaluation methods and knowledge have been shared with countries centered on joint test exercises, and the evaluation reports of the second and third exercises were released at the Vancouver meeting (third meeting) in July 2025. At the 4th meeting in San Diego in December 2025, it was agreed that the network would be named “International Network for Advanced AI Measurement, Evaluation and Science,” and Japan is deepening international cooperation by introducing evaluation tools for AI safety assessment. In February 2026, the India Conference (5th) was held in conjunction with the India AI Impact Summit, and the document “Consensus and Challenges on AI Evaluation” was released as an agreement document for the network. In Japan, an organized session (OS-42: Safety Measures for Large-Scale Models) was held at the National Conference of the Japan Society for Artificial Intelligence (JSAI2025) in May 2025, in which J-AISI co-organized, and J-AISI reported on the International Network of AI Safety Institutes and joint test exercises.

Related Links:
https://aisi.go.jp/activity/activity_international/250211/

(February 2025, 2nd Joint Test Exercise Report: Blog Post)
https://aisi.go.jp/assets/pdf/International-Network-of-AI-Safety-Institutes-Joint-Testing-Exercise-Improving-Methodologies-for-AI-Model-Evaluations-Across-Global-Languages.pdf
(February 2025, 2nd Joint Test Exercise Report: Summary Version)
https://sgaisi.sg/wp-api/wp-content/uploads/2025/06/Improving-Methodologies-for-LLM-Evaluations-Across-Global-Languages-Evaluation-Report-1.pdf
(June 2025, 2nd Joint Test Exercise Report: Detailed Version)
https://aisi.go.jp/activity/activity_international/250718/
(July 2025, 2nd and 3rd Joint Test Exercise Report: Article)
https://sgaisi.sg/wp-api/wp-content/uploads/2025/07/International-Joint-Testing-Exercise_3JT-Blogpost.pdf
(July 2025, 3rd Joint Test Exercise Report: Summary Version)
https://sgaisi.sg/wp-api/wp-content/uploads/2025/07/International-Joint-Testing-Exercise_3JT-Eval-Report-v2.pdf
(July 2025, 3rd Joint Test Exercise Report: Detailed Version)
https://www.gov.uk/government/news/efforts-to-share-best-practices-on-ai-measurement-and-evaluations-driven-forward-through-the-international-network-for-advanced-ai-measurement-evalua
(December 2025, UK: Network Announcements)
https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions
(February 2026, UK: Consensus and Challenges on AI Assessment)

(3) Conclusion of a Memorandum of Cooperation (MOC) with U.S.-based Anthropic

On Wednesday, October 29, 2025, the Japan AI Safety Institute (J-AISI, Director: Akiko Murakami) signed a memorandum of cooperation (MOC) with Anthropic, an American AI development company (co-founder and CEO Dario Amodei). This initiative serves as a memorandum for cooperation for a trusted AI ecosystem.
https://aisi.go.jp/activity/activity_international/251029/

(4) Singapore AI Safety Red Teaming Challenge

J-AISI continues to participate in the International Red Teaming Challenge in Singapore. In addition to AISI, domestic private security companies were invited to participate in the January 2026 session, gaining technical knowledge on red teaming and promoting international cooperation through exchanges with participating countries. We will return the knowledge gained in the field of practical attacks and evaluations to discussions and initiatives on AI safety in Japan and abroad.

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

4) Expansion into Industry and Society

(1) Release of the Chief AI Officer Guide (Chief AI Officer Guidebook and Practical Manual for Establishing a Chief AI Officer and Implementing AI Governance)

With the growing adoption of generative AI and agentic AI, the use of AI is expanding across all aspects of business operations. At the same time, compound risks related to legal compliance, security, privacy, fairness, and accountability are also increasing. In this environment, effective controls cannot be achieved through departmental efforts alone, and a unified governance and operating framework is increasingly essential to ensure that AI functions “safely and effectively” across the organization.

The “Chief AI Officer Guides” aim to provide standard practical guidance for private sector organizations in establishing and operating a Chief AI Officer (CAIO) function, and to support organizations in balancing value creation through AI (maximizing ROI) and responsible AI use (risk reduction and regulatory compliance). By clarifying the role of the CAIO and presenting an integrated view of organizational design, processes, evaluation and oversight, training, talent, and procurement, the goal is to help organizations steadily advance AI adoption in line with their own context.

The scope encompasses organizational design and roles and responsibilities, operational processes, templates, KPIs, and audit and

reporting mechanisms, spanning the full AI lifecycle from planning through procurement to operations. It does not provide detailed technical specifications for individual algorithms or model architectures; rather, it focuses on “governance, organization, and processes” that enable organizations to manage and continuously improve AI over time.

The Chief AI Officer Guides consist of two documents: (1) the Chief AI Officer Guidebook, intended primarily for the CAIO and the appointing authority; and (2) the Practical Manual for Establishing a Chief AI Officer and Implementing AI Governance, intended primarily for the CAIO and the CAIO’s operational staff, senior executives considering the appointment of a CAIO, and relevant stakeholders such as legal and compliance, the IT function, data management functions, human resources, and business divisions. It is hoped that the CAIO Guides will serve as a common language to help organizations move beyond one-off AI deployments and build trust and competitiveness over the long term. (Released in March 2026)
https://aisi.go.jp/output/output_framework/260317/

(2) Establishment of the J-AISI Business Demonstration Working Group, Activities, and Results

In March 2025, the Business Demonstration Working Group (WG) on AI Safety was established with the aim of widely disseminating AI safety assessment activities to the general public and promoting the use of AI. This year, four sub-working groups (Robotics SWG, Healthcare SWG, Data Quality SWG, and Conformity Assessment SWG) have been formed under the Business Demonstration Working Group, with more than 30 organizations participating.

The vision paper was released in June 2025 with the goal of serving as an agenda and guideline for the realization of an AI society where trust and innovation are compatible. Each SWG proceeded with its activities by selecting use cases and conducting evaluation and demonstration, and reported the results of the activities at the first and second half report meetings.

At the international forum “Hiroshima Global Forum for Trustworthy AI” hosted by the Cabinet Office and AISI, the activities of the Business

Demonstration Working Group were introduced.

As an AI safety promotion and awareness activity in the field of robotics, a video was released on YouTube and exhibited at the International Robot Exhibition.

https://aisi.go.jp/lp01/ (Japanese Only)

(Project Demonstration Results Report LP)

https://aisi.go.jp/output/output_information/250630/

(Release of vision paper)

https://aisi.go.jp/activity/activity_information/251003/ (Japanese Only)

(2025 J-AISI Project Demonstration Working Group Report Meeting)

https://aisi.go.jp/activity/activity_information/260311/ (Japanese Only)

(Hiroshima Global Forum for Trustworthy AI)

https://aisi.go.jp/activity/activity_international/260130/

(J-AISI Robotics SWG introduction video)

https://www.youtube.com/watch?v=pKGVWA8F3p8&t=47s (Japanese Only)

(3) AI Safety Activity Map and Terminology Challenge (Visualization of the Whole Picture of Activities)

Based on the organization and benchmarking of international documents, the “Activity Map on AI Safety” was published in February 2025, which comprehensively summarizes the activities necessary to realize AI safety, so that it can provide a bird's-eye view of the relationships between areas and activities that are often overlooked. An automated analysis framework for policy documents using published maps is scheduled to be published as an academic paper in March 2026.

https://aisi.go.jp/output/output_information/250207_1/

(4) Project “Building AI Safety Benchmarks with All Japan/One Team” (Creation of an evaluation method based on cross-organizational collaboration)

With the aim of concretizing AI safety evaluation procedures from the developer's point of view, and building and proposing them as benchmarks, J-AISI led the launch of a discussion framework and held a kickoff in October 2025. For each classification such as usage scenarios,

contents, and countermeasures, subcommittees are studying design policies and preparing sample data throughout FY2025.
https://aisi.go.jp/activity/activity_information/251009/ (Japanese Only)

## 1.3 Other AISI-related activities and achievements

### 1) AI Security Opinion Exchange Meeting

In order to clarify the reality of AI security, which is still not a consistent definition, we have set up a forum to listen to the opinions of experts from industry, government, and academia, focusing on the following three points.

- Strengthening AI systems from the cybersecurity perspective
- Understanding vulnerabilities and emerging threats in AI-embedded social systems
- Vulnerability assessment of social systems embedded with AI

Starting with February 2025, it was held four times in the form of an opinion exchange meeting: July, October, and January 2026.

Since the second meeting, not only experts but also 11 IT-related industry groups have joined, and even more extensive and active discussions have been developed.

### 2) Questionnaire survey report on explaining how AI works, analyzes, and uses

This survey has been conducted since 2024 from the perspective of fixed-point observation of the actual state of AI utilization in Japan. The questionnaire asked respondents about issues such as the appropriateness of explanations provided to users by AI system providers and AI business managers, as well as the state of accountability regarding the performance, reliability, and proper use of AI systems.

The results of the survey are available on the IPA Technical Watch page on the IPA website.
https://www.ipa.go.jp/security/reports/technicalwatch/20251023.html
(Japanese Only)

### 3) Research on the integration of AI governance and cybersecurity risk management in the United States

J-AISI conducted a survey with a U.S. research firm on the integration of AI

governance and cybersecurity risk management in the U.S., an advanced AI market centered on generative AI. Specifically, we focused on interviews and case studies, researching and analyzing AI use cases, risks, and best practices, with a focus on the financial sector, where AI is becoming more utilized.

The results of this survey were widely disseminated to AISI-related ministries and agencies through the study session.

4) Resource cross-walk (standardization of cross-referencing procedures between resources)

In order to enable discussions on interoperability, including consistency and complementarity, for various resources (guides, evaluation datasets, etc.) that will increase with the spread of generative AI, we will standardize the cross-walk procedure and propose it as a “resource cross-walk”. This is an effort to standardize the methodology for the activities of the Guide Cross Walk, which J-AISI has also been working on, and plans to publish a discussion paper outlining it in March 2026.

5) International Activities

In addition to exchanging views, both in person and online, with AI safety stakeholders in various countries, including through the International Network for Advanced AI Measurement, Evaluation and Science, J-AISI has also given lectures and served as instructors for training programs.

Main meeting partners

- Exchange of views with ASEAN, the United Kingdom, the European Commission, Australia, Canada, Kenya, the Republic of Korea, France, the United States, Malaysia, and Malta
- Exchange of opinions with overseas AI-related companies
- JICA Japan International Cooperation Agency Training Program
- A Joint Training Program organized with the ASEAN-Japan Cybersecurity Capacity Building Centre (AJCCBC) and the Japan International Cooperation Agency (JICA)

6) India AI Impact Summit

J-AISI participated in the India AI Impact Summit and related events held from February 16 to 20, 2026. In addition to speaking at the main summit and

at “AI Safety Connect Day,” a side event, J-AISI exchanged views with relevant organizations and enhanced Japan’s presence in the international AI safety community in a strategic and multifaceted manner. At the heads-of-institute meeting, participants also exchanged practical views on the future operation of the International Network for Advanced AI Measurement, Evaluation and Science.

7) Domestic Promotion Activities

As part of its domestic AI safety awareness activities, J-AISI has given lectures and participated in panel discussions at various forums and events, including the Japan Business Federation (Keidanren), the AI Governance Association, the Tokyo Metropolitan SME Intellectual Property Symposium, the LLM Symposium, and the AI Quality Management Symposium.

## 1.4 AISI's Structure

### 1) Overall

In cooperation with the Cabinet Office, a framework for collaboration among 12 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 Cybersecurity Office, 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 Health, Labour and Welfare, Ministry of Agriculture, Forestry and Fisheries, Ministry of Economy, Trade and Industry, Ministry of Land, Infrastructure, Transport and Tourism, 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, Japan (IPA)

### 2) Japan AI Safety Institute Partnership

In order to effectively promote AISI's activities, J-AISI has concluded partnership agreements with five related organizations with expertise in AI safety: National Institute of Information and Communications Technology (NICT), RIKEN, National Institute of Informatics (NII), National Institute of Advanced Industrial Science and Technology (AIST), and Information-technology Promotion Agency, Japan (IPA). J-AISI works with these organizations to conduct research and surveys and to disseminate information both domestically and internationally.

In September 2025, J-AISI released its AI safety efforts at each of its partner organizations.

https://aisi.go.jp/activity/activity_j-aisi_partnership/250929/

3) AISI Secretariat

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

- Strategy and Planning Team
  - Coordination with government agencies
  - Coordination with private sector activities
  - International coordination
  - Project management
  - General planning
- Framework Team
  - Survey
  - Guide development
  - Framework international coordination
- Technology Team
  - Survey
  - Activity map development
  - Technical guide development
  - Technical verification
  - International technical coordination
- Security Team
  - Survey
  - International security technical coordination
- Standards Team
  - Survey of trends in international standards (including activities in

ISO/IEC JTC 1/SC 42 (Artificial intelligence))

- Standard international coordination

■ Utilization Team

- Planning and coordination of the Business Demonstration Working Group
- Investigation of AI use cases related to AI safety

As of March 2026, J-AISI has approximately 30 staff members, including both full-time and part-time personnel. Although J-AISI is small in scale compared with organizatios such as the U.S. AISI (CAISI), with approximately 60 staff members, the UK AISI, with approximately 200 staff members, and the EU AI Office, with approximately 60 staff members, it brings together experts and undertakes cutting-edge activities in AI safety as the scope of AI safety continues to expand.

## 1.5 Challenges due to the special situation of AISI

J-AISI is organized as a technical team specializing in AI safety, but from the user's point of view, it is easier to understand and integrate initiatives that include the use of AI safety rather than just digging deeper. In addition, Appendix 3, “Proper data training,” for AI developers in the AI Guidelines for Business states that developers should verify that the data does not contain personal data, confidential information, rights including copyrights, or legally protected interests. This highlights the need to address data quality. In “Consensus and Challenges on AI Evaluation,” released in connection with the India AI Impact Summit in February 2026, one area of consensus was brought up that evaluation should be conducted with a clear purpose. At the same time, international discussions have progressed on both policy and technical aspects, including the examination of actual use cases and approaches, as well as issues concerning the future direction of AI measurement and evaluation. From an AI safety perspective, J-AISI needs to work with relevant stakeholders to consider how to address areas that cannot be solved by technology alone.

J-AISI is actively working to secure human resources in the AI field in collaboration with domestic partners as it is required to respond to a wide range of stakeholders, including ministries and agencies, engineers, and overseas.

However, in the current situation where there is a shortage of AI human resources both domestically and internationally, restrictions on the treatment of civil servants are also affected, making it a difficult situation.

In addition, there are many related organizations, and since it is a virtual organization, it is required to operate under different information management policies and environments, and the governance system is multi-layered, making it difficult to respond flexibly and promptly.

In addition, the OECD listed 10 risks as possible future AI risks in its "Assessing potential future Artificial Intelligence risks, benefits and policy imperatives" released in November 2024. In addition to technical risks, it points out the risk that "governance mechanisms and organizations will not be able to keep up with the rapidly changing AI reforms."

## 2 Looking Ahead

### 2.1 New factors to consider

In addition to the challenges of 1.5, the following factors should be considered for the future:

1) Laws and Regulations and Basic Plan

The Act on Promotion of Research and Development, and Utilization of AI-related Technology (AI Act) was promulgated on June 4, 2025, and fully implemented on September 1, 2025. In addition, in the Basic Plan for Artificial Intelligence (AI) based on Article 18 of the AI Act (Cabinet decision on December 23, 2025), "AI governance initiative" is proposed as a measure to promote the research and development and utilization of related technologies, and it is described as a fundamental strengthening of AISI, such as immediately expanding the number of personnel to about twice the current level, technical evaluation of AI models, and understanding the actual situation related to the risks posed by AI.

2) Technological Advancements

The development of AI systems is accelerating as the field of AI utilization expands, such as AI agents that are rapidly developing and physical AI implemented in robots, etc. These new AI and the expansion of AI applications may lead to the emergence of new AI safety challenges. It may

also change the scope, perspective, and evaluation methodology of AI safety.

3) AI safety agencies in other countries

I. Initiatives of International Network for Advanced AI Measurement, Evaluation and Science

Based on discussions at the Paris AI Action Summit in February 2025 and the AI Impact Summit in India in February 2026, the International Network for Advanced AI Measurement, Evaluation and Science is evolving from exchanging information to sharing concrete best practices. There are also moves within the International Network for Advanced AI Measurement, Evaluation and Science to explore bilateral technical cooperation.

II. Initiatives by International Organizations and Other Organizations

Discussions on AI are also underway in the G7, as well as international organizations such as the OECD and the United Nations, and the content, promotion system, and specific methods of its considerations are expected to have an impact both domestically and internationally. The debate mainly revolves around themes such as the ethical use of AI, how to regulate it, and strengthening international cooperation, which could have a significant impact on the policies and systems of various countries, including Japan.

4) Collaboration with private companies

AISIs in the UK and US are working with private companies to evaluate AI models and develop evaluation tools. These assessment tools are also designed with a strategic perspective to consider their practicality across industry and society, and to address long-term impacts and risks. While closely monitoring the cooperation between the U.S. government and private companies, Japan also needs to explore cooperation to conduct AI safety assessments.

## 2.2 Future Directions

Based on the current issues surrounding AI and new factors to consider, we will take necessary measures through the following initiatives with the aim of making Japan one of the world's leading countries in the development and utilization of AI, with J-AISI having the information and technology necessary to provide safe, secure, and reliable AI.

1) Establish evaluation metrics for AI safety

Regarding the guide to ensure the safety of AI, we will revise existing guides to respond to the latest evolving AI and create guides that are specific to the issues specific to individual industrial fields. In addition, we will develop benchmarks for each industry and enhance data sets and promote the formulation of Japan's first international standard.

2) Develop in-house capabilities for AI evaluation

We will start building an evaluation environment that enables third-party evaluation of AI, and lead to the evaluation of frontier AI abuse using the AI evaluation environment we have built. In addition, as a response to cyberattacks and defenses by AI, we will collect and evaluate information related to AI incidents and share information with relevant government agencies.

3) Collect, analyze, and provide AI-related information

In addition to providing industry-specific AI guidance and technical advice on system design through public-private collaboration, we will collect and analyze a wide range of information on AI introduction cases and AI incident cases in areas with a high impact (infrastructure fields such as power grids and mobility).

In addition, technical support for investigations into Article 16 of the AI Act (deepfakes, copyright infringement, etc.) and technical information to prevent the distribution of high-risk AI products will be provided to related parties.

4) Leading international cooperation

In order to contribute to the activities of 1) ~ 3) above, we will strengthen cooperation through the International Network for Advanced AI Measurement, Evaluation and Science and actively participate in the sharing of best practices. In addition, we will also provide technical cooperation for capacity building in emerging countries and collaborate with international organizations to ensure that J-AISI can lead the international response to the evaluation of AI of concern.