Japan AI Safety Institute (J-AISI)

August 01 2025

AISI Japan AI Safety Institute

Background and Overview of J-AISI

Establishment of AISI in Japan

AISI Japan AI Safety Institute

Following the **Hiroshima AI Process** and the UK-hosted **AI Safety Summit**, the Japan AI Safety Institute (J-AISI) was established in Feb. 2024.

May 2023	December 2023	February 2024	May 2024	July 2024	March 2025
Agreed to the Hiroshima AI Process	Agreement on "Hiroshima AI Process Comprehensive Policy Framework"	Japan AI Safety Institute (J-AISI) was established	Hiroshima AI Process Friends Group was established	GPAI Tokyo Expert Support Center was established	HAIP reporting framework was started

June 2025
AI act

Japan participates in and contributes to global initiatives.

Hiroshima AI Process
AI Safety Summit 2023

Integrated Innovation Strategy 2024

AISI Japan AI Safety Institute

In the “Integrated Innovation Strategy 2024”,
J-AISI is designated as **the central institution for AI Safety in Japan.**

- The Integrated Innovation Strategy 2024 is the fourth annual strategy that is positioned as the implementation plan for the 6th Science, Technology, and Innovation Basic Plan by the Cabinet Office.

Three strengthening measures of the Integrated Innovation Strategy 2024

1. **Integrated strategy for key technologies**
2. **Strengthening cooperation from a global perspective**
3. **Enhancing competitiveness and ensuring safety and security in AI field**

① **AI innovation and AI accelerated innovation** (Strengthening R&D capabilities, promoting the use of AI, upgrading infrastructure, etc.)
② **Ensuring AI safety and security** (Governance, safety considerations, countermeasures against false information and misinformation, intellectual property, etc.)
③ **Promoting international cooperation and collaboration** (International cooperation based on the outcomes of the Hiroshima AI Process, etc.)

Integrated Innovation Strategy 2025

AISI Japan AI Safety Institute

Strategy 2024

Designated J-AISI as the central institution for AI Safety in Japan.

One of the three pillars of strengthening measures is AI

Enhancing competitiveness and ensuring safety and security in AI field

① **AI innovation and AI accelerated innovation**
(Strengthening R&D capabilities, promoting the use of AI, upgrading infrastructure, etc.)

② **Ensuring AI safety and security**
(Governance, safety considerations, countermeasures against false information and misinformation, intellectual property, etc.)

③ **Promoting international cooperation and collaboration**
(International cooperation based on the outcomes of the Hiroshima AI Process, etc.)

Strategy 2025

(1) Strategic Promotion of Advanced Science and Technology

① Strategic Promotion of Key Areas

- **Balancing "the promotion of innovation through AI" and "risk management"**
- **Advancing AI research and development**
- **Development and promotion of shared use of AI-related facilities**
- **Promoting the utilization of AI**
- **Ensuring responsible use of AI**
- **Securing AI-related talent and promotion of education**
- **Research and studies related to AI**
- **Promoting international cooperation in the AI field**

Role and Scope of J-AISI

AISI Japan AI Safety Institute

J-AISI's role is to support public and private sector initiatives to promote the safe and secure use of AI.

Role

- Primarily plays three roles.

Support the government

- Investigating AI safety, examination of evaluation methods, and creating standard.

Hub of AI Safety in Japan

- Collecting the latest industry-academia initiatives.
- Promoting collaboration among related entities.
- Collaborating with international AI safety institutions.

Collaboration with AI Safety-related organizations

- Collaborate with national research institutes.
- Promote partnerships

Building a framework that enables AI developers and users to **correctly recognize AI-related risks**

+

Building a framework that enables **the implementation of necessary measures**, such as ensuring governance, **throughout the entire lifecycle**

⇔ **Domestic & international related organizations**

Achieving a framework that balances "Promotion of Innovation" and "risk mitigation throughout the lifecycle."

Scope

- Setting the scope flexibly, while considering global trends regarding AI-related issues.

Social Impact	Governance	AI System	Contents	Data

Initiatives for ensuring AI safety

AISI Japan AI Safety Institute

J-AISI undertakes **safety evaluations**, **implementation methods**, as well as **international collaboration**.

1. **Research on evaluation of safety standards and criteria.**
   - Investigation of safety standards, check tools, disinformation countermeasures, and AI and cybersecurity.
   - Consideration of safety standards and guidelines.
   - Consideration of a testbed environment for AI related to the above.
2. **Study of methodologies for implementing safety evaluations.**
3. **International collaboration** with relevant organizations in other countries.

AISI Japan AI Safety Institute

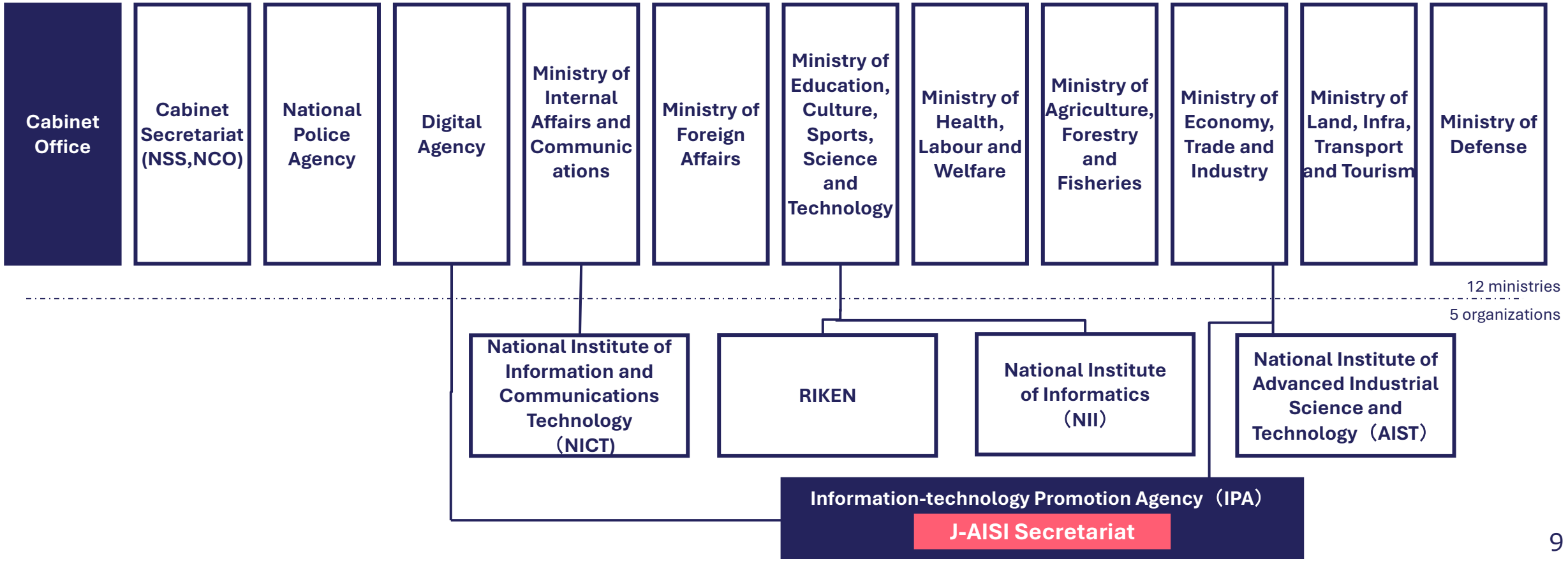
Organizational Structure

Related Government organization and agencies

AISI Japan AI Safety Institute

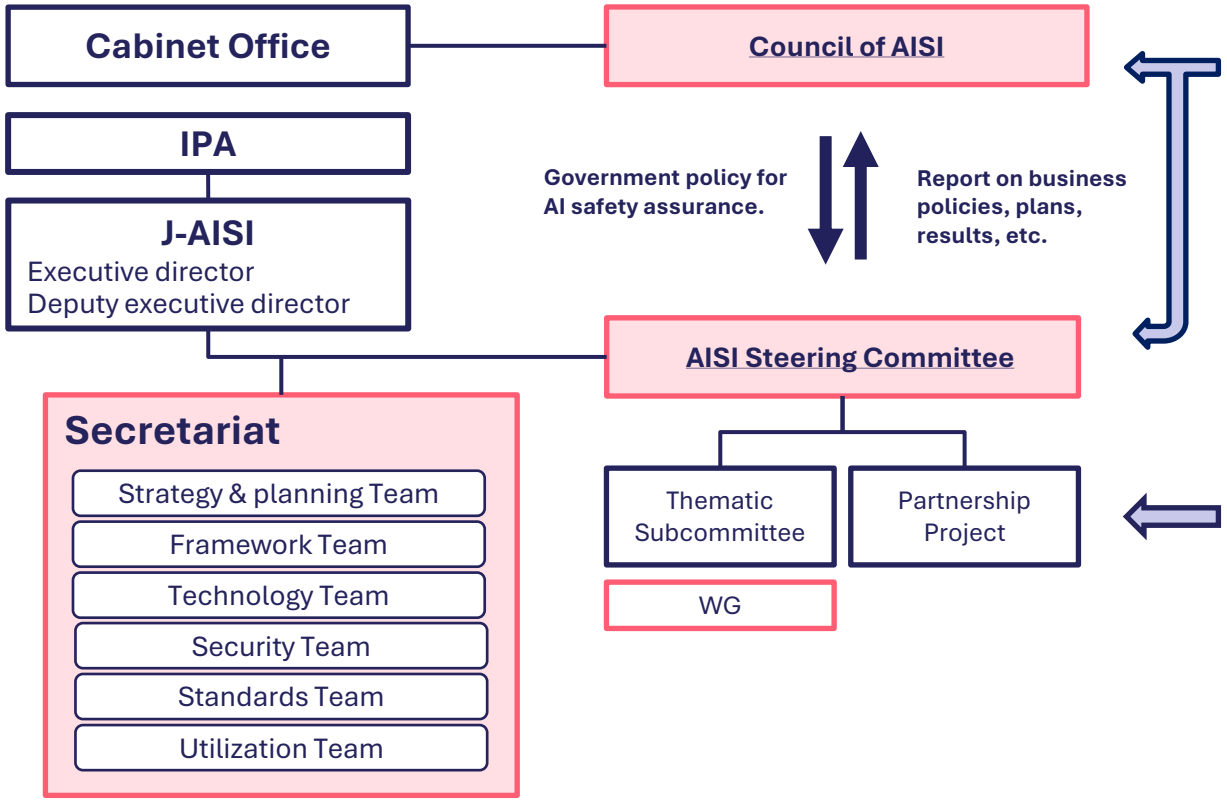
AISI is a government-related organization in which 12 ministries and agencies, along with 5 related organizations, participate cross-sectionally. The secretariat is set within the IPA, under the jurisdiction of the METI* and the Digital Agency.

*METI: Ministry of Economy, Trade and Industry



J-AISI Structure

Government policies reviewed by the Council of AISI, led by the Cabinet Office.
Project policies assessed by the AISI Steering Committee, chaired by the AISI Director.



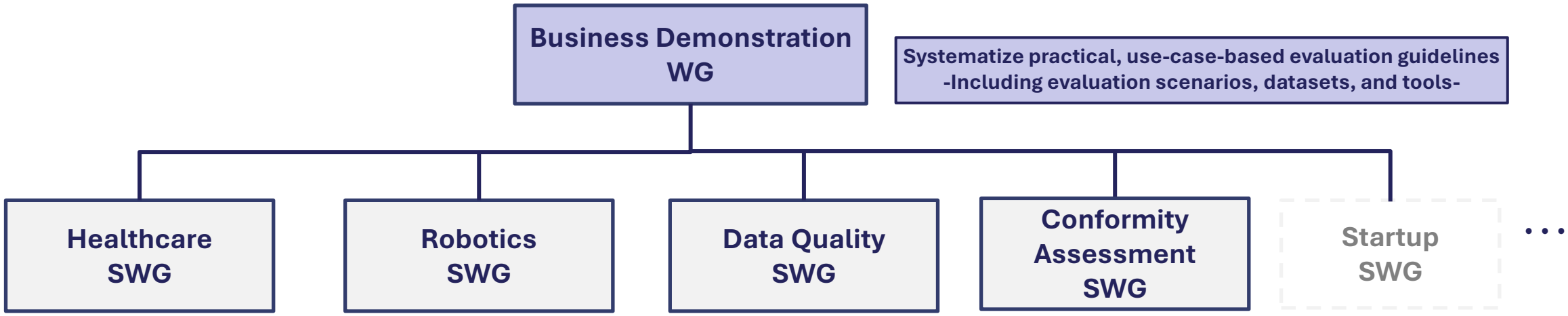
Relevant Ministries and Agencies:
- Cabinet Office (Secretariat of Science, Technology and Innovation Policy)
- National Security Secretariat
- National Cybersecurity Office
- National Police Agency
- Digital Agency
- Ministry of Internal Affairs and Communications
- Ministry of Foreign Affairs
- Ministry 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:
- IT Promotion Agency, Japan (IPA)
- National Institute of Information and Communications Technology
- RIKEN
- National Institute of Informatics
- National Institute of Advanced Industrial Science and Technology

Establishment of WG on AI Safety Evaluation

AISI Japan AI Safety Institute

- The Business Demonstration WG brings together a wide range of stakeholders, primarily from the private sector, to identify use cases and needs from practical perspectives, share business insights, discuss challenges, and organize the findings into a systematic framework.
- The activities of WG aim to establish AI safety evaluation practices that contribute to the social implementation and industrial deployment of AI, with the goal of supporting not only government agencies but also private-sector businesses and engineers.



Japan AI Safety Institute Partnership

AISI Japan AI Safety Institute

To promote AI Safety Initiatives through the cooperation of relevant organizations, J-AISI Partnership was issued in August, 2024.

- To advance initiatives related to AI Safety, it is essential for J-AISI and Domestic organizations to collaborate and jointly address the challenges.

  - The following is an excerpt from the "Japan AI Safety Institute Partnership Agreement.

Article 3 (Activities of the Partnership)

The Japan AI Safety Institute Partnership ("the Partnership") shall, in collaboration with participating organizations stipulated under Article 5, promote the following activities within the scope agreed upon between AISI and the participating organizations in accordance with the provisions of Article 7, Paragraph 2, to effectively advance AISI's initiatives:

1. Joint research and studies conducted by AISI and participating organizations regarding AI safety.
2. Provision of advice by participating organizations concerning activities carried out by AISI.
3. Sharing of information with AISI regarding activities on AI safety conducted by participating organizations.
4. Dissemination of information domestically and internationally, as well as coordination and collaboration with domestic and international relevant organizations, regarding the initiatives outlined in the preceding items by AISI and participating organizations.
5. Other activities incidental to the initiatives outlined in the preceding items.

Executive Team

Executive Director **Akiko Murakami**

1999: IBM Japan, Research Laboratory
2016: IBM Japan, Software Development Laboratory
2021: Sompo Japan Insurance Inc.
Executive Officer, CDaO (Chief Data Officer),
General Manager of the Data-Driven Management Promotion Department [Current]
2024: Executive Director of J-AISI [Current]
2025: Group Chief Data Officer Executive Vice President Sompo Holdings,Inc. [Current]

Deputy Executive Director/
Secretary General

Kenji Hiramoto

1990: NTT DATA Japan Corporation
2008: Executive Advisor to CIO, METI
2012: Chief Strategist(IT), Cabinet Secretariat
2021: Head of Data Strategy(Executive officer), Digital Agency
2023: Director General, IPA Digital Infrastructure Centre [Current]
2024: Deputy Executive Director, Secretary General of J-AISI [Current]

Deputy Executive Director

Suguru Nishimura

2000: Ministry of Posts and Telecommunications
2006: Consul at Japanese Consulate General in Sydney, Australia
2009: Deputy Director, Multilateral Economic Affairs Office,
International Economic Division, Global ICT Strategy Bureau,
MIC (Responsible for APEC, OECD, EPA)
2021: Director, National Healthcare Policy Secretariat, Cabinet Office
2024: Director, Office of the Director-General for Cybersecurity , MIC
2025: Deputy Executive Director of J-AISI [Current]

Secretariat

AISI Japan AI Safety Institute

The Secretariat is composed of the following six teams and includes many seconded personnel from government and private companies.

Strategy & planning Team

- Strategies and planning, Budget management
- PR, Human resource development
- Coordination with domestic and international organizations

Technology Team

- Establishment of evaluation methods for AI safety
- Development of evaluation environments

Standards Team

- Establishment of conformity assessment methods in the AI field
- Consideration of building a domestic framework for practical implementation

Framework Team

- Consideration of an evaluation framework for AI safety
- Coordination to ensure interoperability in AI governance

Security Team

- Research on specific attack methods on AI systems
- Consideration of a classification system for AI security incidents
- Systematization of attacks targeting AI systems

Utilization Team

- Planning and coordination of Business Demonstration Working Group
- Research on AI utilization cases related to AI safety

Activities and Deliverables

Activities and Deliverables for FY2024

AISI Japan AI Safety Institute

		International	J-AISI	Government
		EVENT	DELIVERABLE	
2024	Apr		• **JP-U.S. Crosswalk1**(4/30)	• **AI Guidelines for Business** was published(4/19)
	May	AI Safety Summit, Korea		
	Jun	G7 Summit, Italy		• **Integrated Innovation Strategy 2024** was published(6/4)
	Jul		• Japanese Translation of U.S. AI RMF(7/4)	
	Aug			
	Sep		• **Guide to Evaluation Perspectives**(9/18) • **JP-U.S. Crosswalk2**(9/18) • **Guide to Red Teaming Methodology***(9/25)	
	Oct			
	Nov	International Network of AISIs Convening, USA		
	Dec		• **National Status Report on AI Safety in Japan**(2/5) • **Activity Map on AI Safety**(2/7) • **Data Quality Management Guidebook**(Draft) (2/7) • **Known Attacks and Their Impacts on AI Systems**(3/26) • **AI Safety・Approach book**(3/26) • **Guide to Evaluation Perspectives on AI Safety (v.1.10)**(3/28) • **Guide to Red Teaming Methodology on AI Safety(v.1.10)**(3/31) • **Data Quality Management Guidebook**(3/31)	
2025	Jan			
	Feb	AI Action Summit, France		• Updated on **AI Guidelines for Business** (3/28)
	Mar			

* Red teaming involves identifying and addressing weaknesses in AI systems from an attacker's perspective to maintain or enhance AI safety

Summary of Deliverables

AISI Japan AI Safety Institute

We prioritize our efforts, from technical reviews to human resource development, with the AI guidelines for Business at the center.

Crosswalk
For international interoperability

AI Guidelines for Business
Developed and updated by the MIC* and the METI*

Activity Map on AI Safety
comprehensive overview and prioritization

Guide to Evaluation Perspectives
Evaluation

Guide to Red Teaming Methodology
Red Teaming

Data Quality Management Guidebook
To provide qualified data for AI

Multilingual/ Multicultural
Multilateral challenges

Security Report
Knowledge of security

Digital Skills Standards
Human resource development

*MIC: Ministry of Internal Affairs and Communications
*METI: Ministry of Economy, Trade and Industry

AI Guidelines for Business

AISI Japan AI Safety Institute

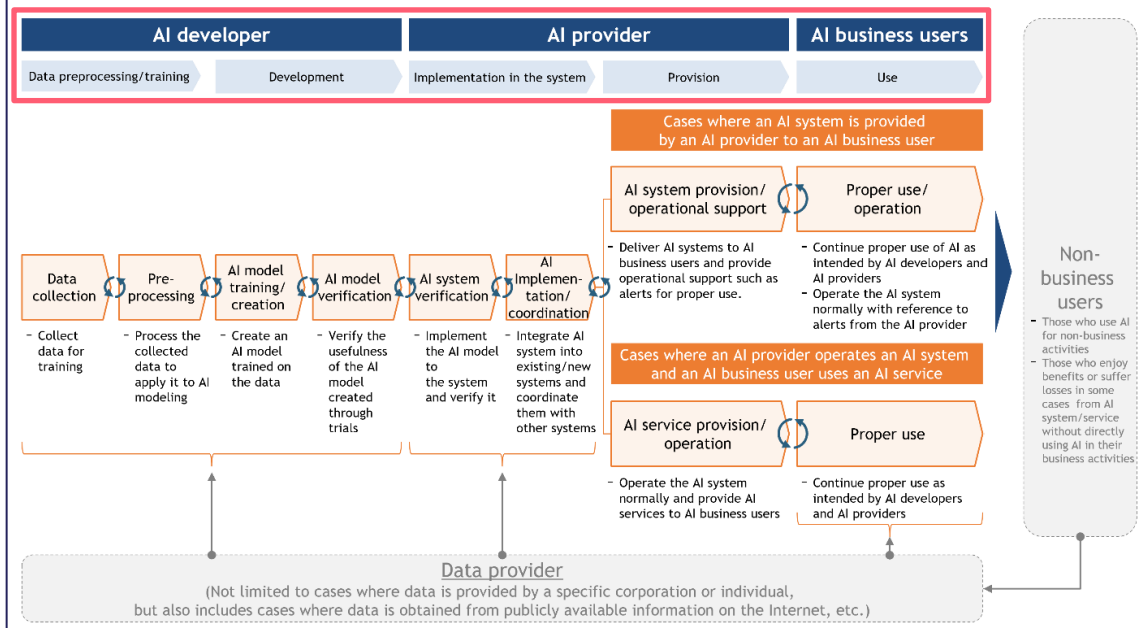
METI* and MIC* have integrated and updated the existing guidelines, and published the AI Guidelines for business (Ver 1.0).

*Updated to Ver 1.1 in March 2025.

- Clarifying the responsibilities of each stakeholder in the process of utilizing AI.
- J-AISI co-hosts the study group for “AI Guidelines for Business” with METI.

Content intended as a guide for businesses

Defining the content by each lifecycle stage and entity



Defining governance Behavioral Goals for private operators

Category	Behavioral Goals ※ Some are further subdivided like「3-1-1」
1. Environmental and risk analysis	1-1 Understanding benefits/risks 1-2 Understanding social acceptance of AI 1-3 Understanding company's AI know-how
2. Goal setting	2-1 Setting AI governance goals
3. System design	3-1 Requiring evaluation of goal deviation and measures to minimize it 3-2 Improving literacy of those in charge of the AI management system 3-3 Enhancing AI management through cooperation between AI business actors and divisions 3-4 Reducing burden related to incidents involving AI Business Users and non-business users through preventive and prompt action
4. Operation	4-1 Ensuring that the operation of AI management system is explainable 4-2 Ensuring that the operation of each AI system is explainable 4-3 Considering proactive disclosure of AI governance practices
5. Evaluation	5-1 Verifying AI management system functions 5-2 Considering opinions of outside stakeholders
6. Environment and risk reanalysis	6-1 Reimplementing Behavioral Goals 1-1 to 1-3 at an appropriate time

*METI: Ministry of Economy, Trade and Industry, *MIC: Ministry of Internal Affairs and Communications

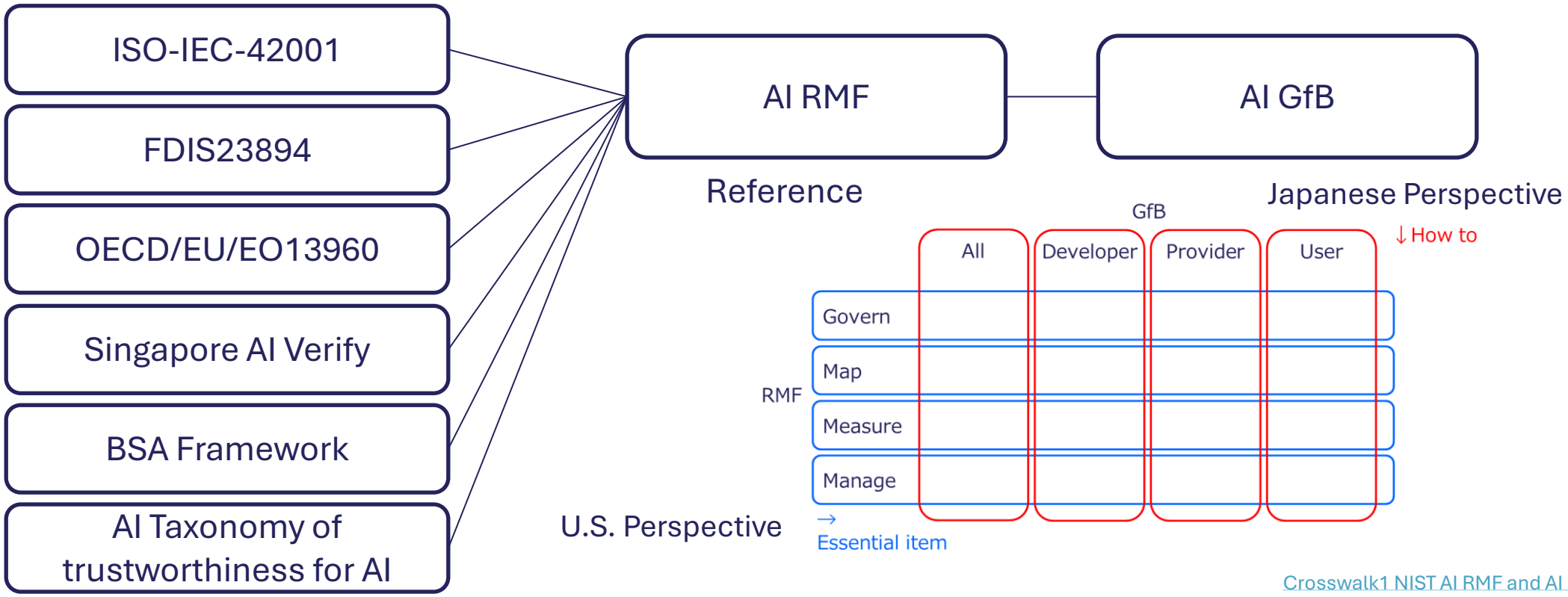
AI Guidelines for Business（METI）, AI Guidelines for Business（MIC）

Japan-U.S. Crosswalk

AISI Japan AI Safety Institute

Confirmation of the interrelationship between the U.S. NIST AI Risk Management Framework and the Japanese AI Guidelines for Business.

- Using the US AI RMF as a reference, it is possible to cross-check with guidelines from other countries.



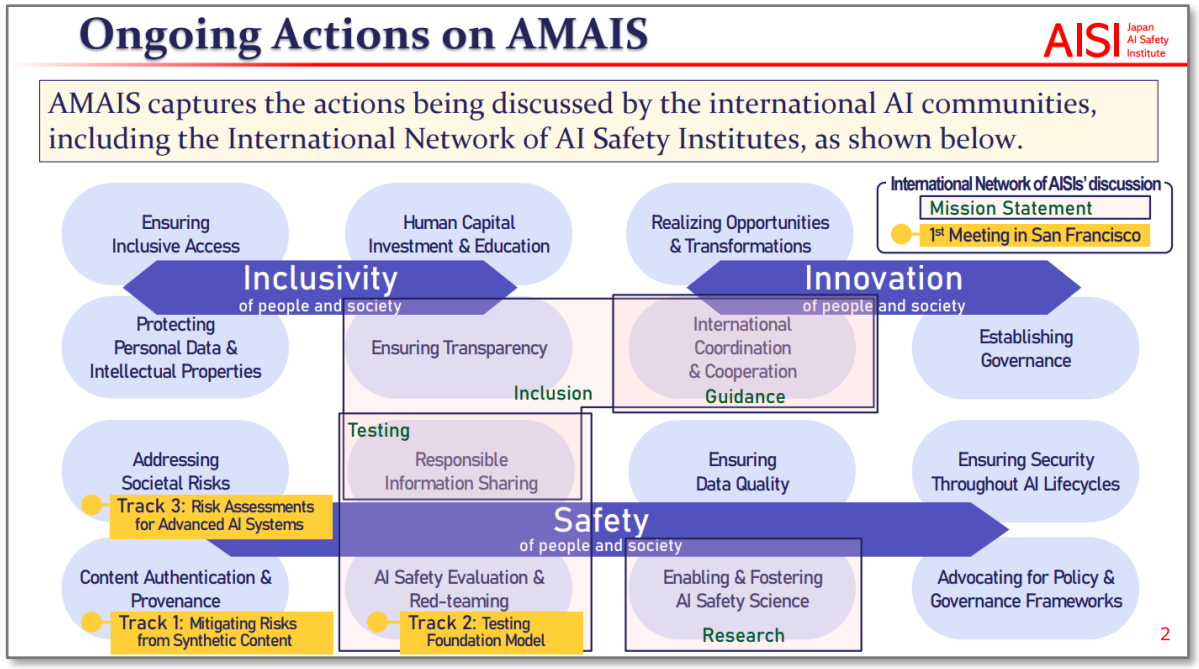
Crosswalk1 NIST AI RMF and AI GfB
Crosswalk2 NIST AI RMF and AI GfB

AMAIS: Activity Map on AI Safety

AISI Japan AI Safety Institute

AMAIS visualizes the overall picture of overlooked areas and correlations between activities as the AI safety efforts rapidly evolve.

- J-AISI has published "AMAIS: Activity Map on AI Safety" as a **discussion paper**.
- AISI is developing a comprehensive activity map and related terminology based on **benchmarks from key literature**.
- This Japan-led initiative is expected to further strengthen the foundation of international cooperation on AI safety and contribute to the realization of a sustainable and reliable AI society.



AMAIS: Activity Map on AI Safety - AISI Japan

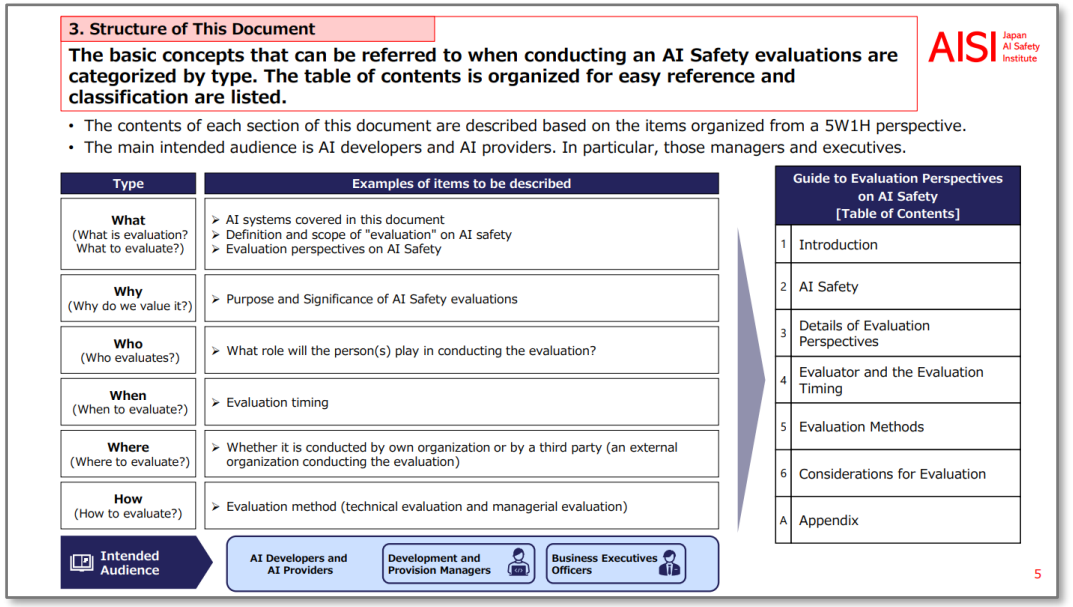
Guide to Evaluation Perspectives on AI Safety

AISI Japan AI Safety Institute

This guide presents **fundamental considerations** that can serve as a reference **when conducting AI safety evaluations.**

- Specifically, this document provides the following:
  - **Perspectives of AI Safety evaluations, examples of risks and evaluation items**
  - **Ideas on who and when it will be conducted**
  - **Summary of evaluation method**

- This guide is a first step towards realizing safe, secure, and reliable AI, and is expected to contribute to maintaining and enhancing safety in future AI development and provision.



3. Structure of This Document

The basic concepts that can be referred to when conducting an AI Safety evaluations are categorized by type. The table of contents is organized for easy reference and classification are listed.

AISI Japan AI Safety Institute

- The contents of each section of this document are described based on the items organized from a 5W1H perspective.
- The main intended audience is AI developers and AI providers. In particular, those managers and executives.

Type	Examples of items to be described
What (What is evaluation? What to evaluate?)	➢ AI systems covered in this document ➢ Definition and scope of "evaluation" on AI safety ➢ Evaluation perspectives on AI Safety
Why (Why do we value it?)	➢ Purpose and Significance of AI Safety evaluations
Who (Who evaluates?)	➢ What role will the person(s) play in conducting the evaluation?
When (When to evaluate?)	➢ Evaluation timing
Where (Where to evaluate?)	➢ Whether it is conducted by own organization or by a third party (an external organization conducting the evaluation)
How (How to evaluate?)	➢ Evaluation method (technical evaluation and managerial evaluation)

Guide to Evaluation Perspectives on AI Safety [Table of Contents]	
1	Introduction
2	AI Safety
3	Details of Evaluation Perspectives
4	Evaluator and the Evaluation Timing
5	Evaluation Methods
6	Considerations for Evaluation
A	Appendix

Intended Audience: AI Developers and AI Providers; Development and Provision Managers; Business Executives Officers

5

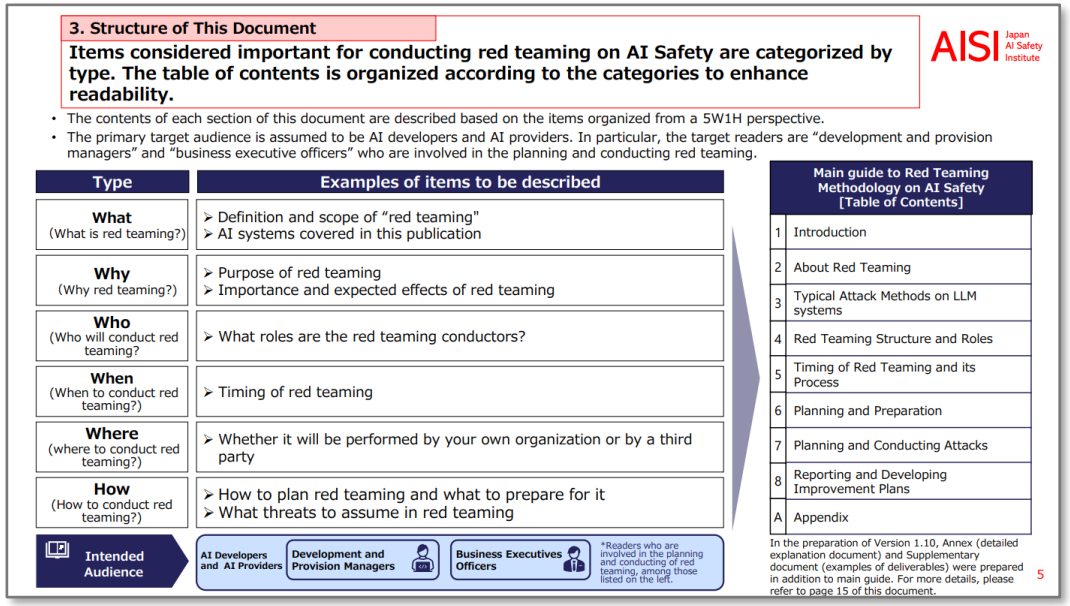
Guide to Evaluation Perspective on AI Safety - AISI Japan

Guide to Red Teaming Methodology on AI Safety

AISI Japan AI Safety Institute

This guide serves as a resource that compiles **fundamental considerations for implementing red teaming methods** in AI safety.

- Specifically, the report provides points to keep in mind regarding the conducting structure, timing, planning, methods, and improvement plans for safety assessments.
- This guide is a first step towards realizing safe, secure, and reliable AI, and is expected to contribute to maintaining and enhancing safety in future AI development and provision.



3. Structure of This Document

Items considered important for conducting red teaming on AI Safety are categorized by type. The table of contents is organized according to the categories to enhance readability.

AISI Japan AI Safety Institute

- The contents of each section of this document are described based on the items organized from a 5W1H perspective.
- The primary target audience is assumed to be AI developers and AI providers. In particular, the target readers are "development and provision managers" and "business executive officers" who are involved in the planning and conducting red teaming.

Type	Examples of items to be described
What (What is red teaming?)	➢ Definition and scope of "red teaming" ➢ AI systems covered in this publication
Why (Why red teaming?)	➢ Purpose of red teaming ➢ Importance and expected effects of red teaming
Who (Who will conduct red teaming?	➢ What roles are the red teaming conductors?
When (When to conduct red teaming?)	➢ Timing of red teaming
Where (where to conduct red teaming?)	➢ Whether it will be performed by your own organization or by a third party
How (How to conduct red teaming?)	➢ How to plan red teaming and what to prepare for it ➢ What threats to assume in red teaming

Intended Audience: AI Developers and AI Providers; Development and Provision Managers; Business Executives Officers

*Readers who are involved in the planning and conducting of red teaming, among those listed on the left.

Main guide to Red Teaming Methodology on AI Safety [Table of Contents]	
1	Introduction
2	About Red Teaming
3	Typical Attack Methods on LLM systems
4	Red Teaming Structure and Roles
5	Timing of Red Teaming and its Process
6	Planning and Preparation
7	Planning and Conducting Attacks
8	Reporting and Developing Improvement Plans
A	Appendix

In the preparation of Version 1.10, Annex (detailed explanation document) and Supplementary document (examples of deliverables) were prepared in addition to main guide. For more details, please refer to page 15 of this document.

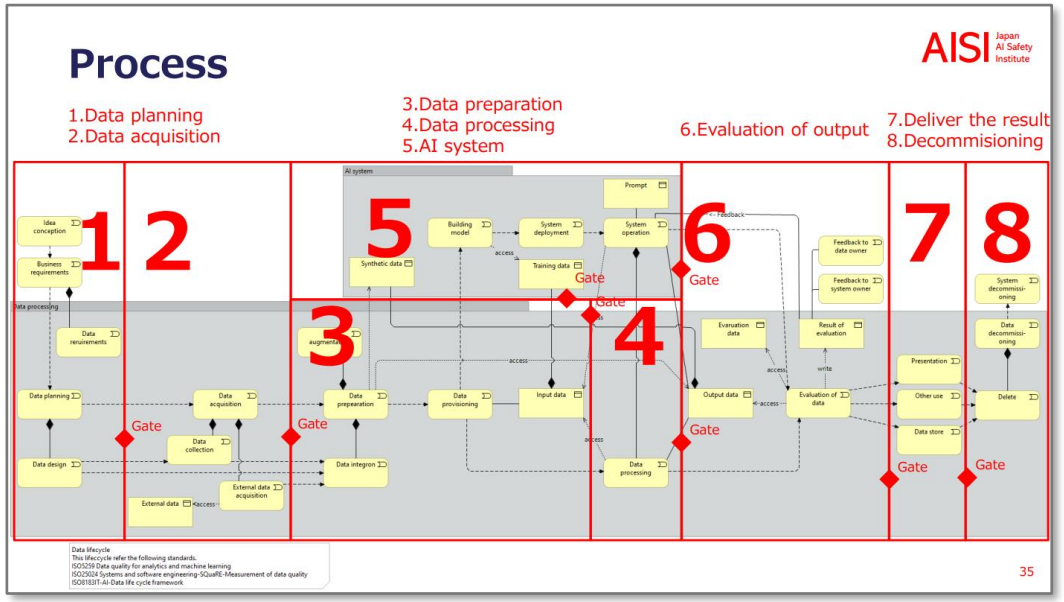
5

Data Quality Management Guidebook

AISI Japan AI Safety Institute

The aim of this guide is to maximize the value of data/AI and **to sustainably ensure data quality**.

- Data quality is the foundation of AI excellence, and contributes to the realization of trust AI. This guide has been organized to help us realize the trust AI society in an appropriate data quality manner and secure the data quality necessary for a data-driven society.
- The English version of this guide is the official version, while Japanese version is a summary.



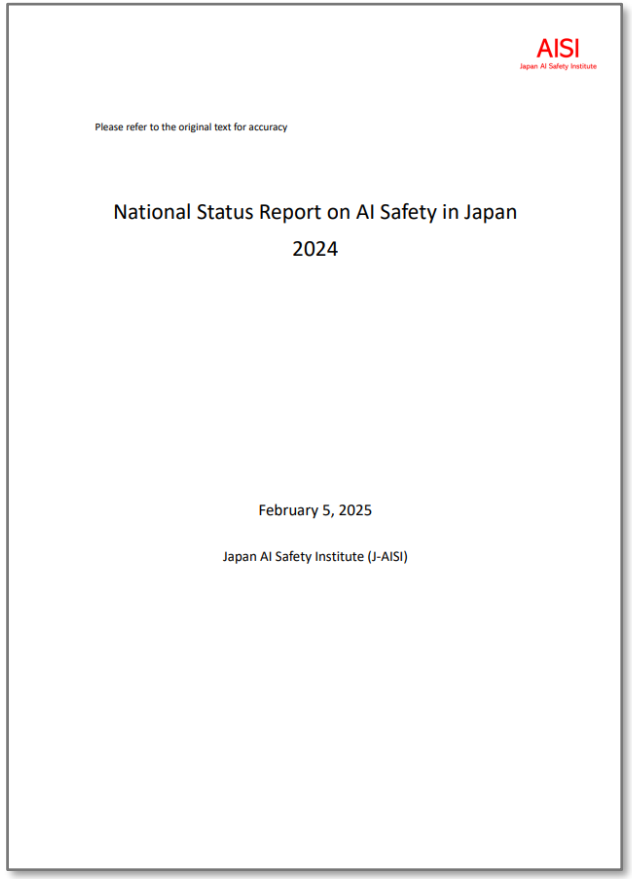
National Status Report on AI Safety in Japan 2024

AISI Japan AI Safety Institute

J-AISI has published the "National Status Report on AI Safety in Japan 2024", covering the **activities of J-AISI.**

- Also compiled related reports in the form of the "Fact Sheet of AI Safety in Japan 2024", a reference document that complements it.

- In the "National Status Report on AI Safety in Japan 2024", we also describe our future initiatives and aims, including collaboration between AISIs and related organizations and the private sector in Japan and overseas to respond to the rapid development of AI.


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)

National Status Report on AI Safety in Japan 2024 - AISI Japan

AISI Japan AI Safety Institute

International Collaboration

Major International meetings

AISI Japan AI Safety Institute

Actively participating in international meetings and engaging in discussions with AI-related businesses and organizations worldwide.

- **AISI-related collaborations**

  - **Stanford University AI Symposium (Stanford/ Apr 16, 2024)**
    - Panel discussion with directors of U.S. and U.K. AISI, and parallel exchange of opinions among countries
  - **AI Seoul Summit (Seoul/ May 21-22, 2024)**
    - High-level roundtable and exchange of views with the U.S., U.K., EU, Canada, Germany, etc.
    - Participation in discussions including Asian and African countries at the concurrent AI Global Forum.
  - **UN Future Summit (UN/ Sep 22, 2024)**
  - **UN Global Compact Leaders Summit 2024 (UN/ Sep 24, 2024)**
  - **AISI International Network Convening (San Francisco/ Nov 10-11, 2024)**
  - **AI Action Summit (Paris/ Feb 6-11, 2025)**
  - **Hiroshima AI Process Friends Group Meeting (Tokyo/ Feb 27-28, 2025)**

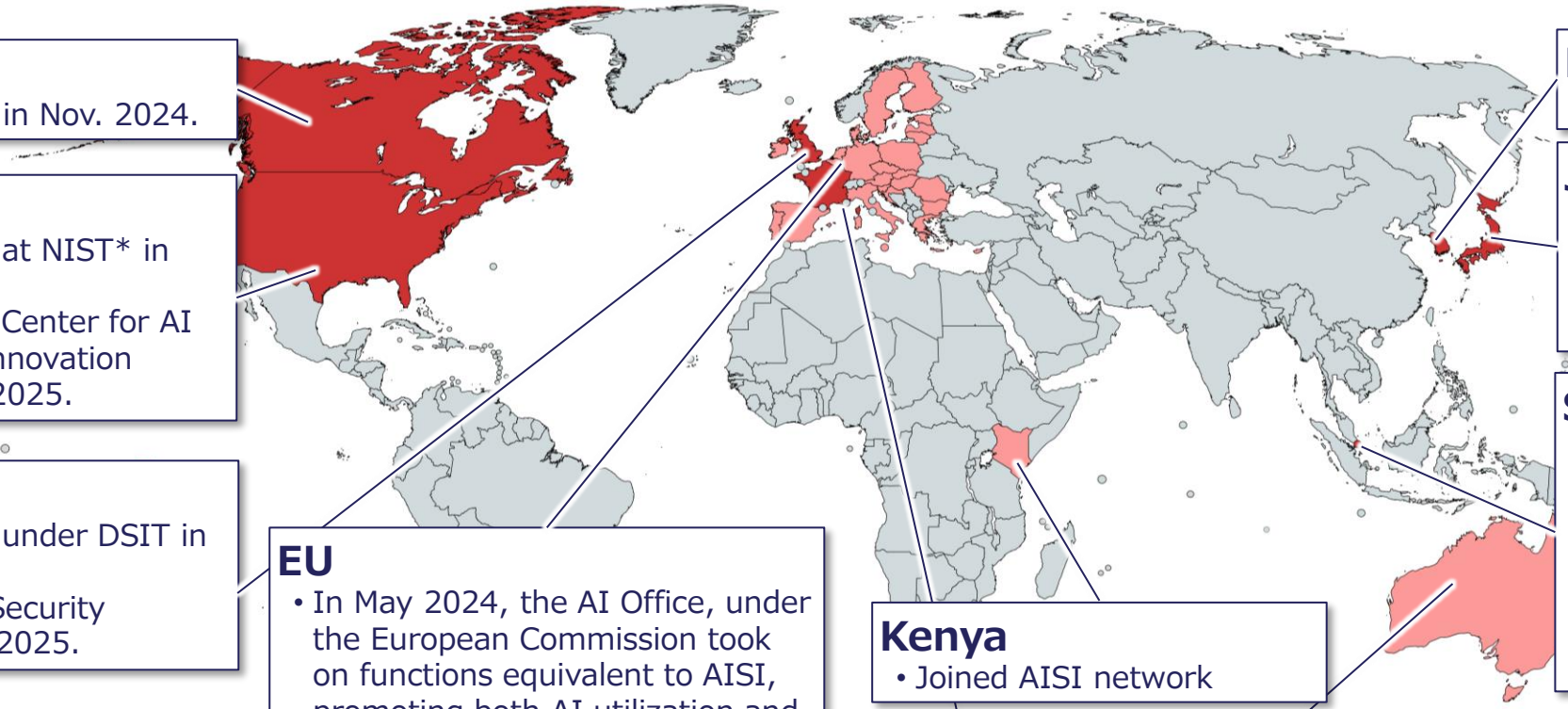
AI Global Forum, held concurrently with the AI Seoul Summit

UN Future Summit

The International Network of AI Safety Institute

AISI Japan AI Safety Institute

Launched under the initiative of U.S. and currently includes participation from 10 leading nations. U.S. serves as the chair, and U.K. acts as the secretariat.



Canada
- Established AISI in Nov. 2024.

US
- Established AISI at NIST* in Feb. 2024.
- Renamed as the Center for AI Standards and Innovation (CAISI) in June 2025.

UK
- Established AISI under DSIT in Nov. 2023.
- Renamed as AI Security Institute in Feb. 2025.

EU
- In May 2024, the AI Office, under the European Commission took on functions equivalent to AISI, promoting both AI utilization and safety. Also responsible for developing and promoting AI Act.

Kenya
- Joined AISI network

Australia
- Joined AISI network.

France
- INESIA (AISI-equivalent institution) was established in February 2025.

Korea
- Established AISI in Nov.2024.

Japan
- Established AISI at IPA* in Feb. 2024.
- Launched as the third AISI following UK and US.

Singapore
- In May 2025, The Digital Trust Centre at NTU* was designated as Singapore AISI.
- Providing safety evaluation test tools aimed at the international standardization of LLMs.

*IPA: Information Promotion
*NTU: Nanyang Technological University

Created with mapchart.net

AISI has been established
AISI-equivalent institutions

*NIST: National Institute of Standard and Technology
*DSIT: Department for Science, Innovation and Technology
*INESIA: l'Institut national pour l'évaluation et la sécurité de l'intelligence artificielle

AISI

Japan AI Safety Institute