Japan AI Safety Institute (J-AISI)

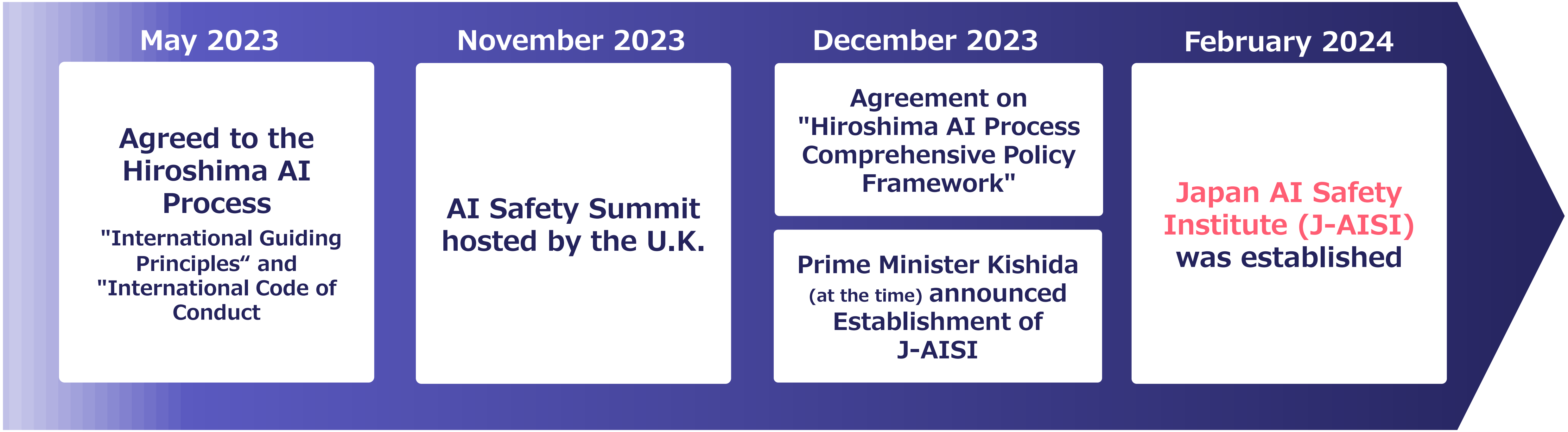
May 1st, 2025

AISI Japan
AI Safety Institute

Background and Overview of J-AISI

Establishment of AISI in Japan

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



May 2023

Agreed to the Hiroshima AI Process

"International Guiding Principles“ and "International Code of Conduct

November 2023

AI Safety Summit hosted by the U.K.

December 2023

Agreement on "Hiroshima AI Process Comprehensive Policy Framework"

Prime Minister Kishida (at the time) announced Establishment of J-AISI

February 2024

Japan AI Safety Institute (J-AISI) was established

Hiroshima AI Process
AI Safety Summit 2023
AI Strategy Council

Integrated Innovation Strategy 2024

AISI Japan AI Safety Institute

In the "Integrated Innovation Strategy 2024",
J-AISI is defined 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 collaboration 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.)

Role and Scope of J-AISI

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

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

1. **Research and evaluation of safety standards and criteria.**
   - Investigation of safety standards, check tools, disinformation countermeasures, and AI and cybersecurity.
   - Development 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. (e.g. AI Safety Institute in the U.K. and the U.S.)

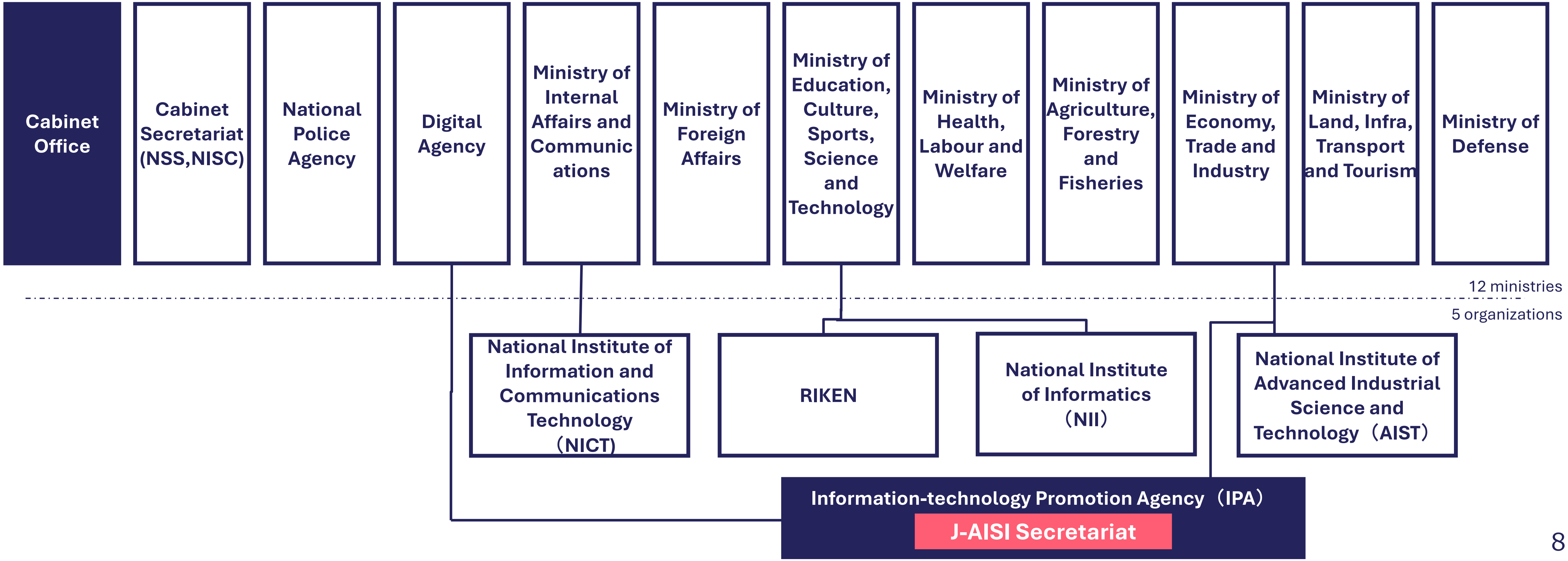
Organizational Structure

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



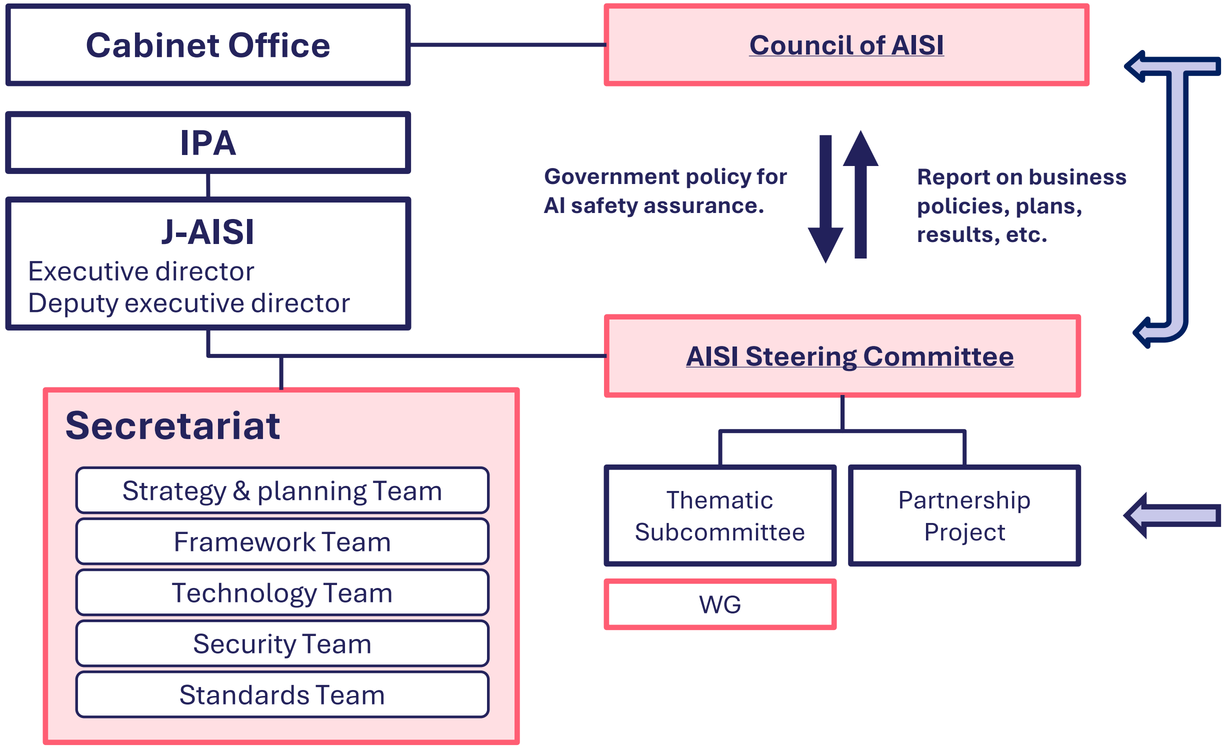
Cabinet Office
Cabinet Secretariat (NSS,NISC)
National Police Agency
Digital Agency
Ministry of Internal Affairs and Communications
Ministry of Foreign Affairs
Ministry of Education, Culture, Sports, Science and Technology
Ministry of Health, Labour and Welfare
Ministry of Agriculture, Forestry and Fisheries
Ministry of Economy, Trade and Industry
Ministry of Land, Infra, Transport and Tourism
Ministry of Defense
12 ministries
5 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）
J-AISI Secretariat

J-AISI Structure

AISI Japan AI Safety Institute

Government policies reviewed by the AISI Liaison Meeting, 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 Center of Incident readiness and Strategy for Cybersecurity
- 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

Executive Team

Executive Director **Akiko Murakami**

1999: Joined IBM Japan, Research Laboratory

2016: Joined IBM Japan, Software Development Laboratory

2021: Joined Sompo Japan Insurance Inc.
Executive Officer, CDaO (Chief Data Officer),
General Manager of the Data-Driven Management Promotion Department [Current]

2025: Group Chief Data Officer Executive Vice President Sompo Holdings,Inc. [Current]

Deputy Executive Director/
Secretary General

Kenji Hiramoto

1990: Joined NTT DATA Corporation

2008: CIO Advisor, METI

2012: Senior Advisor to the Government CIO, Cabinet Secretariat

2021: Director of Data Strategy, Digital Agency

2023: Director, IPA Digital Infrastructure Centre [Current]

2024: Deputy Director, Secretary General of J-AISI [Current]

Deputy Executive Director

Hideyuki Teraoka

1999: Joined the Ministry of Posts and Telecommunications

2007: Ministry of Internal Affairs and Communications,
Telecommunications Bureau

2023: Cabinet Secretariat, Cabinet Cyber Security Center

2024: Deputy Director of J-AISI [Current]

Secretariat

The Secretariat is composed of the following five 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

Activities and Deliverables

Activities and Deliverables for FY2024

		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			
2025	Jan			
	Feb	AI Action Summit, France	- Published **Activity Map on AI Safety**(2/7) - Published **Data Quality Management Guidebook**(Draft) (2/7) - Published **National Status Report on AI Safety in Japan 2024**(2/7)	
	Mar			- Updated on **AI Guidelines for Business** (3/28)

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

Summary of Deliverables

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

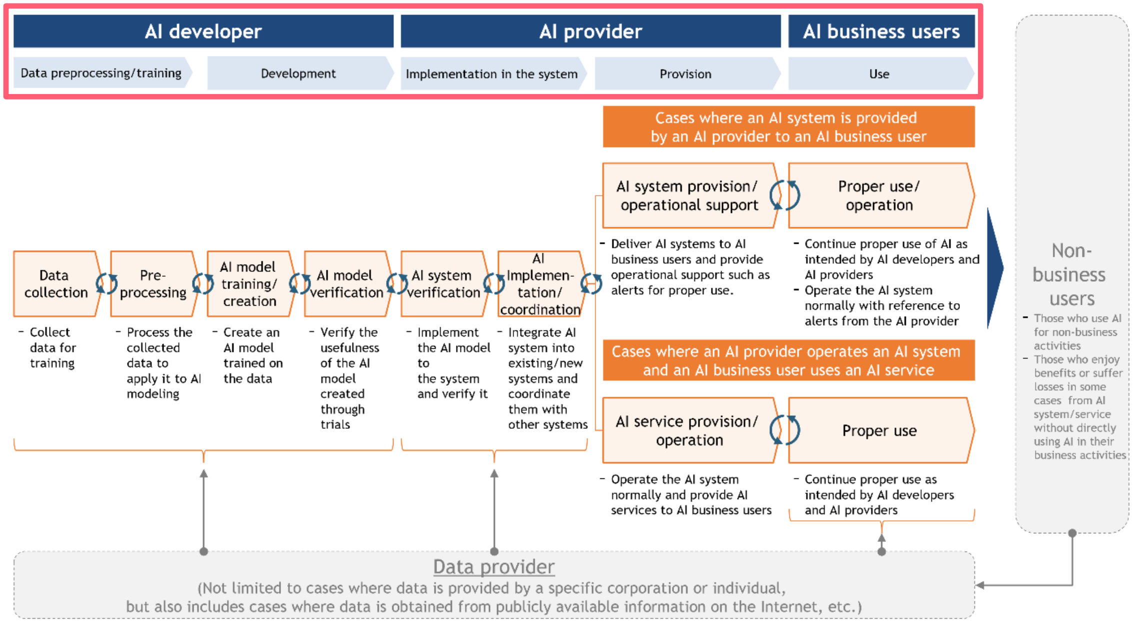
METI* and MIC* have integrated and updated the existing guidelines, and published the AI Guidelines for business (Ver 1.0).

*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**


AI developer
AI provider
AI business users
Data preprocessing/training
Development
Implementation in the system
Provision
Use
Data collection
Pre-processing
AI model training/creation
AI model verification
AI system verification
AI Implementation/coordination
Cases where an AI system is provided by an AI provider to an AI business user
AI system provision/operational support
Proper use/operation
Cases where an AI provider operates an AI system and an AI business user uses an AI service
AI service provision/operation
Proper use
Non-business users
Data provider

- Collect data for training
- Process the collected data to apply it to AI modeling
- Create an AI model trained on the data
- Verify the usefulness of the AI model created through trials
- Implement the AI model to the system and verify it
- Integrate AI system into existing/new systems and coordinate them with other systems
- Deliver AI systems to AI business users and provide operational support such as alerts for proper use.
- Continue proper use of AI as intended by AI developers and AI providers
- Operate the AI system normally with reference to alerts from the AI provider
- Operate the AI system normally and provide AI services to AI business users
- Continue proper use as intended by AI developers and AI providers
- Those who use AI for non-business activities
- Those who enjoy benefits or suffer losses in some cases from AI system/service without directly using AI in their business activities

(Not limited to cases where data is provided by a specific corporation or individual, but also includes cases where data is obtained from publicly available information on the Internet, etc.)

Defining **governance Behavioral Goals** for private operators

※ Some are further subdivided like「3-1-1」

Category	Behavioral Goals
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

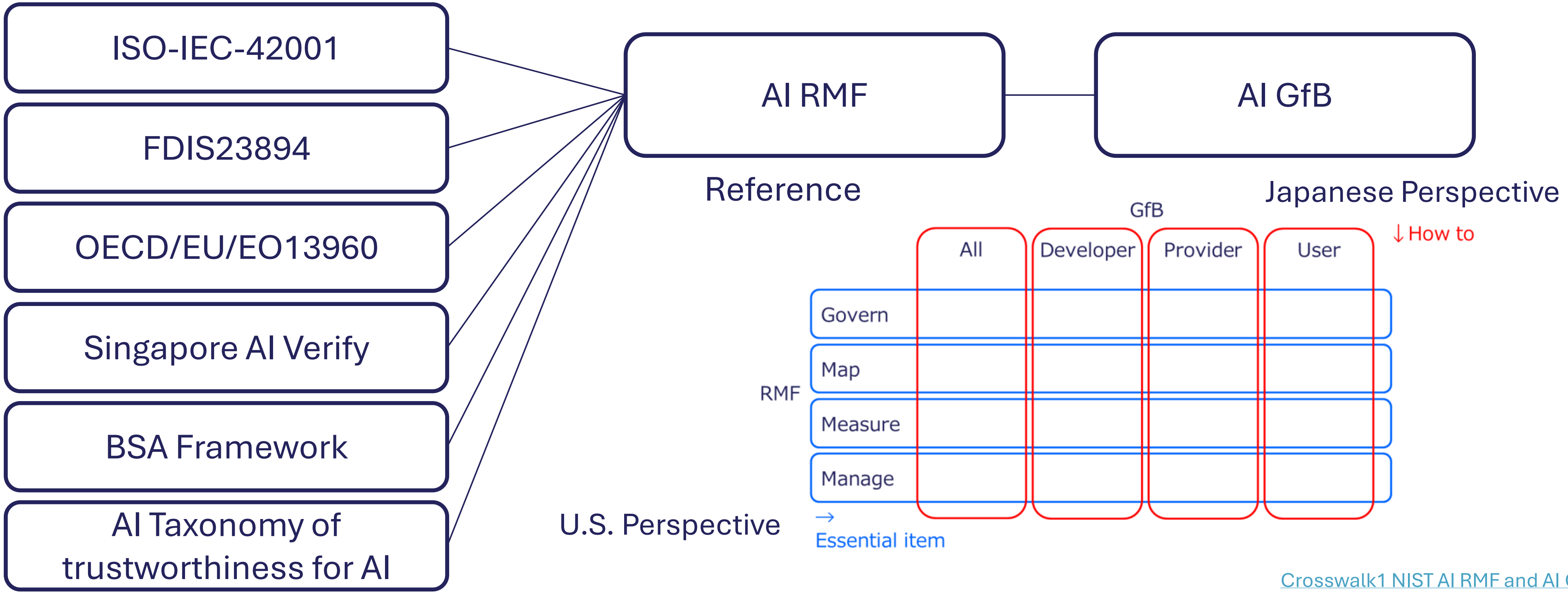
*METI: Ministry of Economy, Trade and Industry, *MIC: Ministry of Internal Affairs and Communications

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

Japan-U.S. Crosswalk

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.



ISO-IEC-42001

FDIS23894

OECD/EU/EO13960

Singapore AI Verify

BSA Framework

AI Taxonomy of trustworthiness for AI

AI RMF

Reference

U.S. Perspective

AI GfB

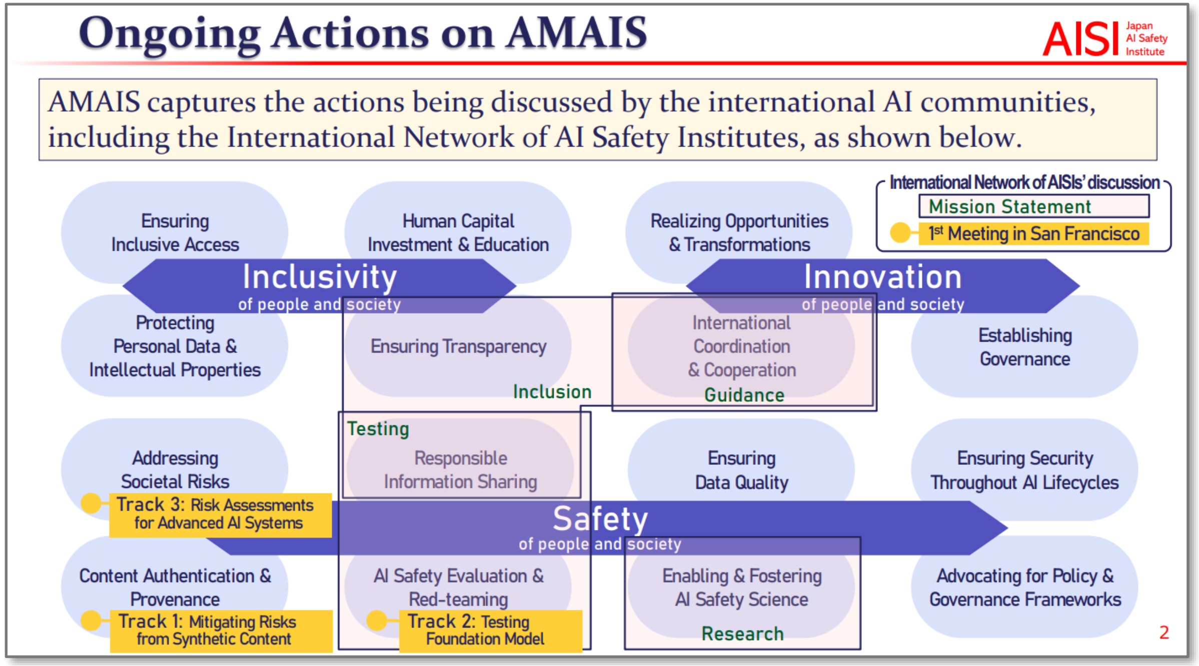
Japanese Perspective

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

AMAIS: Activity Map on AI Safety

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.



Ongoing Actions on AMAIS

AMAIS captures the actions being discussed by the international AI communities, including the International Network of AI Safety Institutes, as shown below.

AISI Japan AI Safety Institute
International Network of AISIs' discussion
Mission Statement
1st Meeting in San Francisco
Ensuring Inclusive Access
Human Capital Investment & Education
Realizing Opportunities & Transformations
Inclusivity
of people and society
Innovation
of people and society
Protecting Personal Data & Intellectual Properties
Ensuring Transparency
Inclusion
International Coordination & Cooperation
Guidance
Establishing Governance
Testing
Addressing Societal Risks
Responsible Information Sharing
Ensuring Data Quality
Ensuring Security Throughout AI Lifecycles
Track 3: Risk Assessments for Advanced AI Systems
Safety
of people and society
Content Authentication & Provenance
AI Safety Evaluation & Red-teaming
Enabling & Fostering AI Safety Science
Research
Advocating for Policy & Governance Frameworks
Track 1: Mitigating Risks from Synthetic Content
Track 2: Testing Foundation Model
2

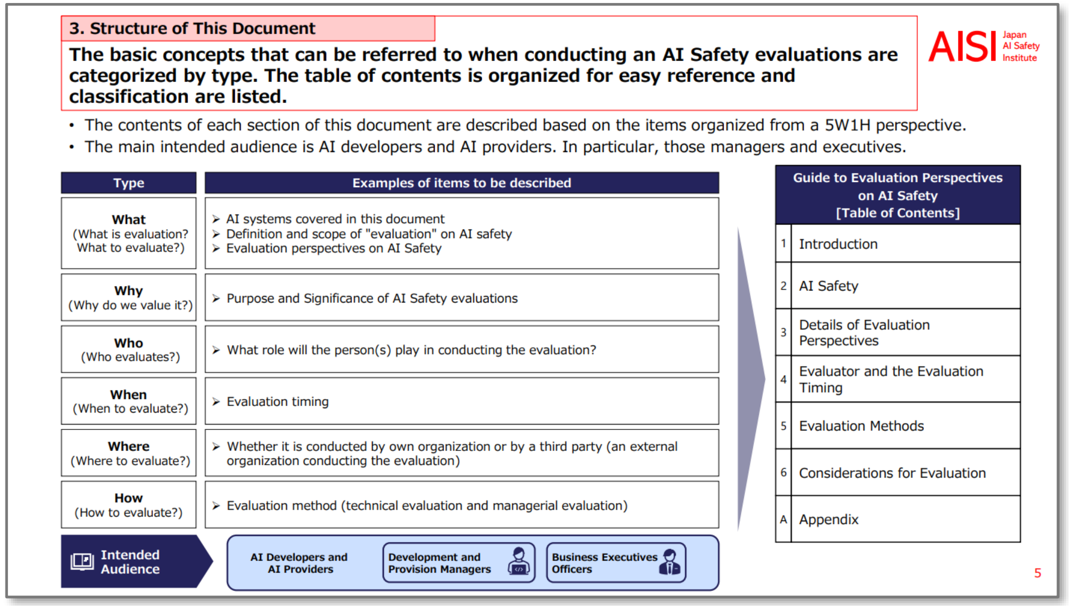
AMAIS: Activity Map on AI Safety - AISI Japan

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)
Intended Audience	**AI Developers and AI Providers** · **Development and Provision Managers** · **Business Executives Officers**

Guide to Evaluation Perspectives on AI Safety [Table of Contents]

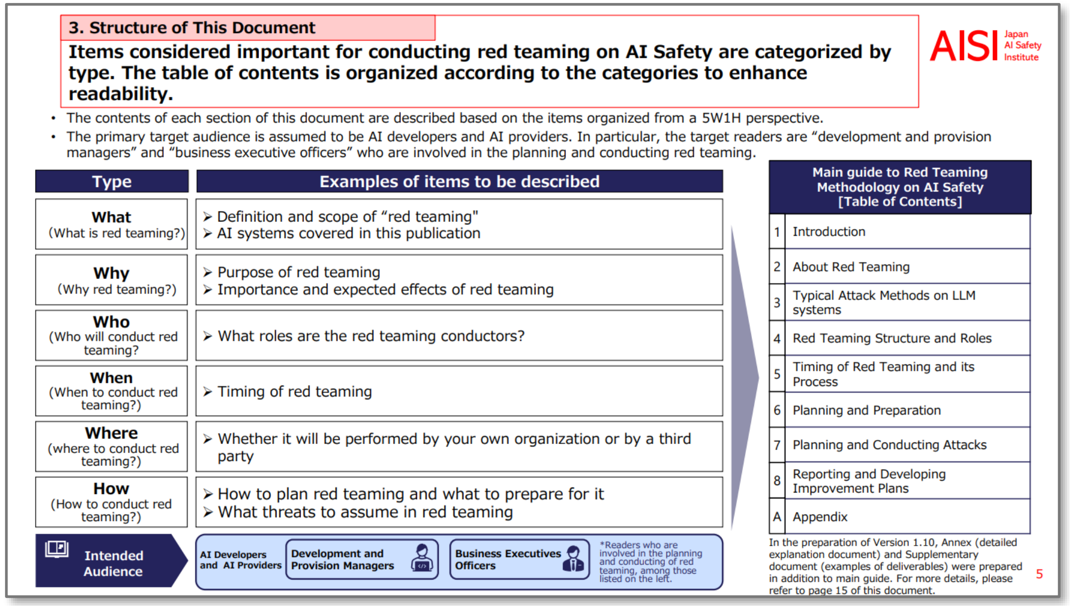
5

Guide to Red Teaming Methodology on AI Safety

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.

- 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.

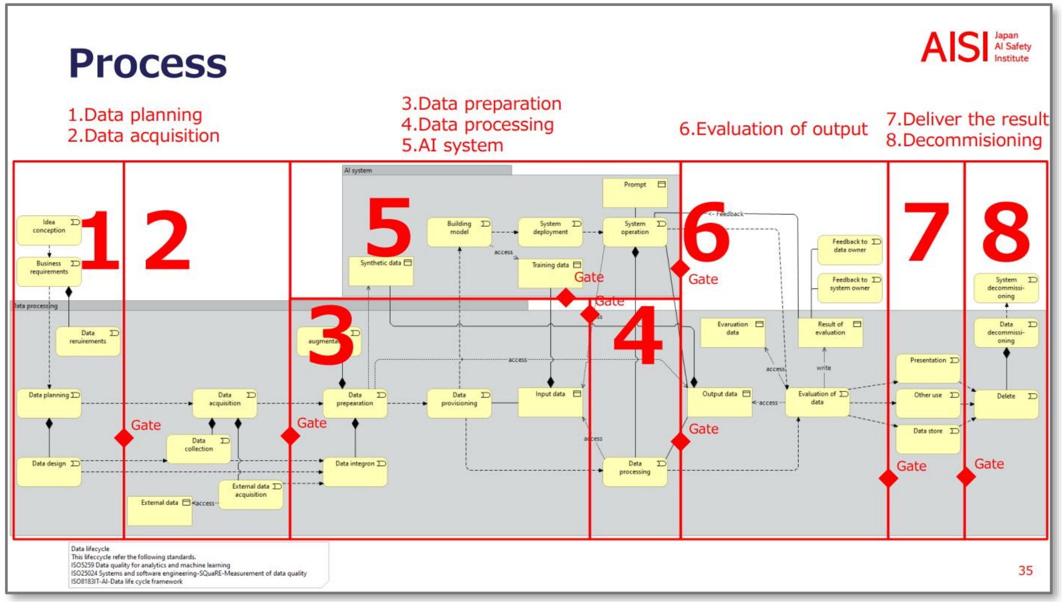
5

Data Quality Management Guidebook

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.



Process

1.Data planning
2.Data acquisition

3.Data preparation
4.Data processing
5.AI system

6.Evaluation of output

7.Deliver the result
8.Decommisioning

Data lifecycle
This lifeccycle refer the following standards.
ISO5259 Data quality for analytics and machine learning
ISO25024 Systems and software engineering-SQuaRE-Measurement of data quality
ISO8183IT-AI-Data life cycle framework

35

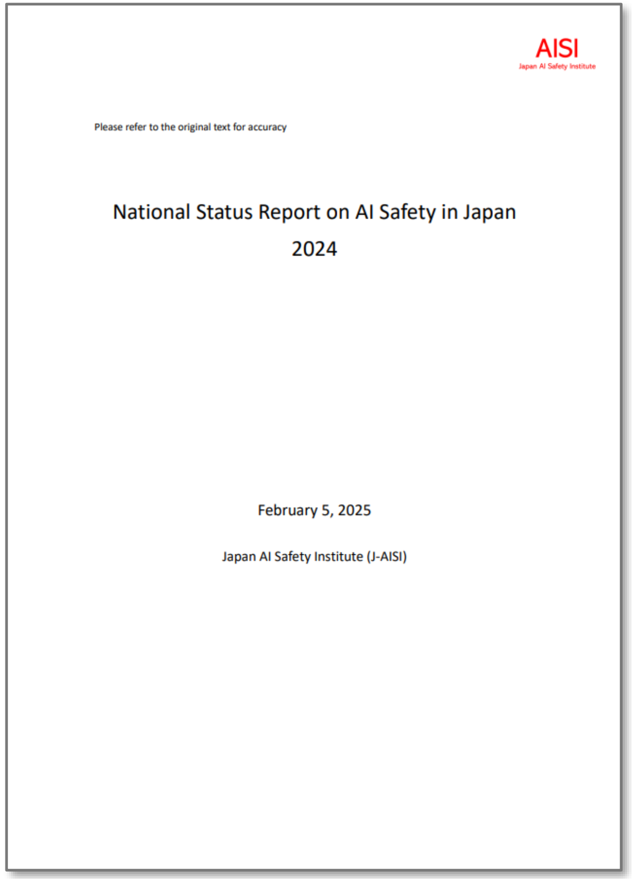
National Status Report on AI Safety in Japan 2024

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

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
22 MAY 2024
Hosted by the Republic of Korea
Governance to Ensure Innovation,

AI Global Forum, held concurrently with the AI Seoul Summit

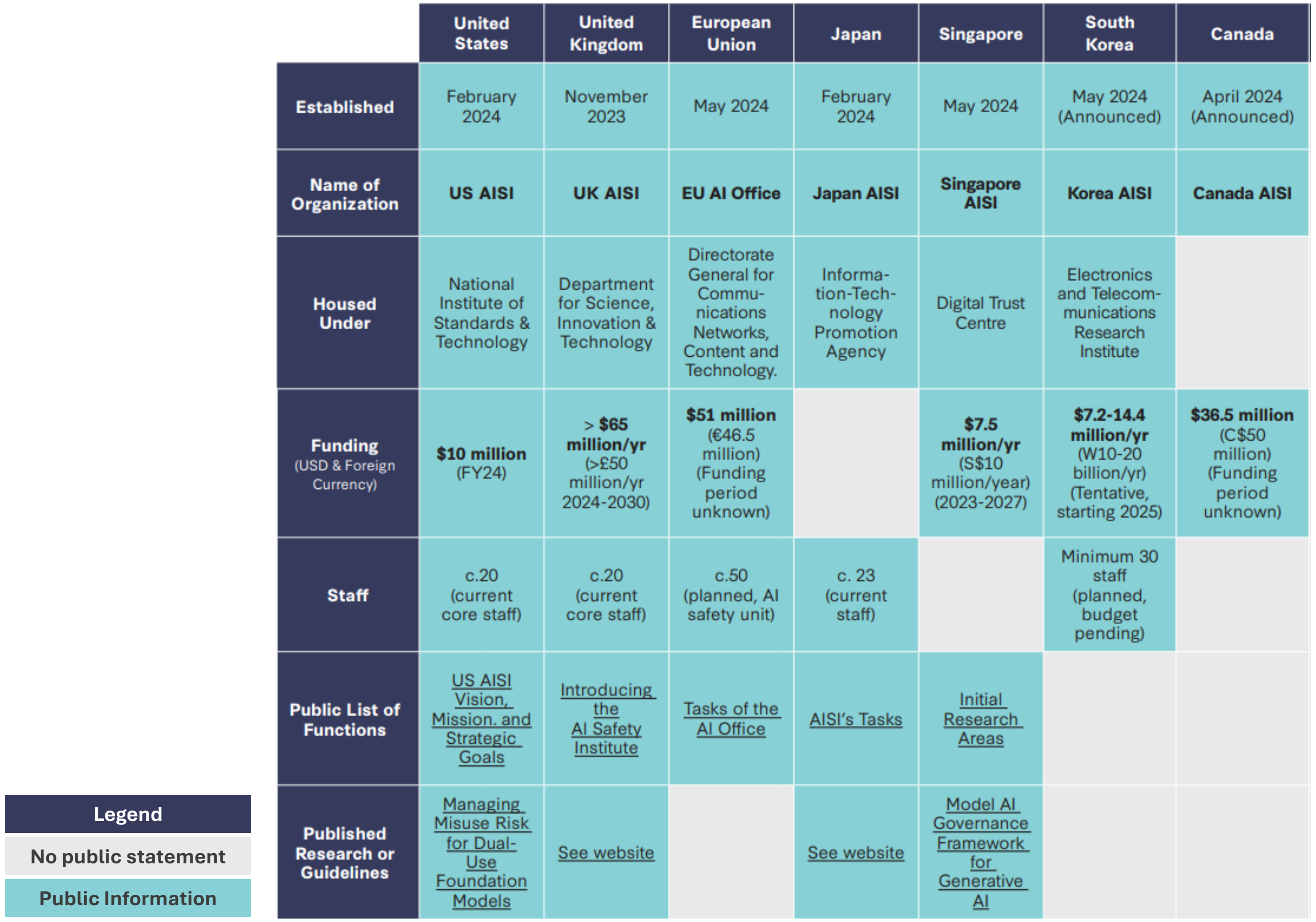
UN Future Summit

Organizational Overview of AISI International Network Members

AISI Japan AI Safety Institute

The status of AISI establishments in various countries is as follows.

*Report by CSIS in October 2024.



	United States	United Kingdom	European Union	Japan	Singapore	South Korea	Canada
Established	February 2024	November 2023	May 2024	February 2024	May 2024	May 2024 (Announced)	April 2024 (Announced)
Name of Organization	**US AISI**	**UK AISI**	**EU AI Office**	**Japan AISI**	**Singapore AISI**	**Korea AISI**	**Canada AISI**
Housed Under	National Institute of Standards & Technology	Department for Science, Innovation & Technology	Directorate General for Communications Networks, Content and Technology.	Information-Technology Promotion Agency	Digital Trust Centre	Electronics and Telecommunications Research Institute	
Funding (USD & Foreign Currency)	**$10 million** (FY24)	> **$65 million/yr** (>£50 million/yr 2024-2030)	**$51 million** (€46.5 million) (Funding period unknown)		**$7.5 million/yr** (S$10 million/year) (2023-2027)	**$7.2-14.4 million/yr** (W10-20 billion/yr) (Tentative, starting 2025)	**$36.5 million** (C$50 million) (Funding period unknown)
Staff	c.20 (current core staff)	c.20 (current core staff)	c.50 (planned, AI safety unit)	c. 23 (current staff)		Minimum 30 staff (planned, budget pending)	
Public List of Functions	US AISI Vision, Mission, and Strategic Goals	Introducing the AI Safety Institute	Tasks of the AI Office	AISI's Tasks	Initial Research Areas		
Published Research or Guidelines	Managing Misuse Risk for Dual-Use Foundation Models	See website		See website	Model AI Governance Framework for Generative AI		

Legend
No public statement
Public Information

The AI Safety Institute International Network, CSIS

AISI

Japan AI Safety Institute