

Overview of the Japan AI Safety Institute (J-AISI)

2024-12-01

AI Strategy

- ◆ AI safety is an essential requirement for an environment that encourages innovation.

① Accelerating AI Innovation and Innovation by AI

- Strengthening R&D Capabilities (Including Data Improvement)
- Promotion of AI Utilization
- Upgrading of Infrastructure
- Development and Securing of Human Resources

② Ensuring **AI Safety**

- Governance and Systems
- Consideration of **AI safety**
- Countermeasures to Dis/Mis-information)
- Intellectual Property Rights

③ Promotion of International Cooperation and Collaboration

① Accelerating AI Innovation and Innovation by AI

- ◆ Strengthening R&D Capabilities (Including Data Improvement)
 - Training data in Japanese and Use cases
 - Support to the start-up projects
 - Advanced AI technology
 - AI for science
 - AI & Robotics
- ◆ Promotion of AI Utilization
 - Use in the government and public services
 - Clarification of institutional operation (e.g., personal information)
- ◆ Upgrading of Infrastructure
 - Computing resources and 5G/6G network
 - Semiconductor
 - Carbon neutral energy
- ◆ Development and Securing of Human Resources
 - Educational resource
 - Support for the next generation researchers

② Ensuring AI Safety

- ◆ Governance and Systems
 - AI guideline for business
 - Policy of rule making
- ◆ Consideration of **AI safety**
 - **J-AISI**
 - R&D for AI safety
- ◆ Countermeasures to Dis/Mis-information
 - Prevention technologies and rules
 - Evaluation methodology
- ◆ Intellectual Property Rights
 - Research the international activities

J-AISI(Japan AI Safety Institute)

- ◆ October 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 Announced Establishment of J-AISI
- ◆ **February 14, 2024**
 - Japan AI Safety Institute (J-AISI) was established

♦ Objectives

- **J-AISI supports public and private sector efforts.**
 - The public and private sectors need to work together to ensure that all parties involved in the development and use of AI are properly aware of the risks of AI. Governance also needs to be ensured throughout the lifecycle. Then, the safe and secure use of AI will be promoted.
 - Need to promote innovation and mitigate risks in the lifecycle, in those efforts.

♦ Principles

- **J-AISI's activities will be harmonized with related organizations in Japan and internationally.**
 - Response to rapidly and globally advancing technologies.

◆ Role


- **J-AISI supports the government** by conducting surveys on AI safety, examining evaluation methods, and creating standards.
- **As a hub for AI safety in Japan**, J-AISI will consolidate the latest information in industry & academia, and promote collaboration among related companies and organizations.
 - J-AISI is not an R&D organization.
- **Collaborate with AI safety-related organizations.**

◆ Scope

- **Set the scope flexibly** in the following AI related issues, while considering **global trends**.
 - Social Impact
 - Governance
 - AI System
 - Contents
 - Data

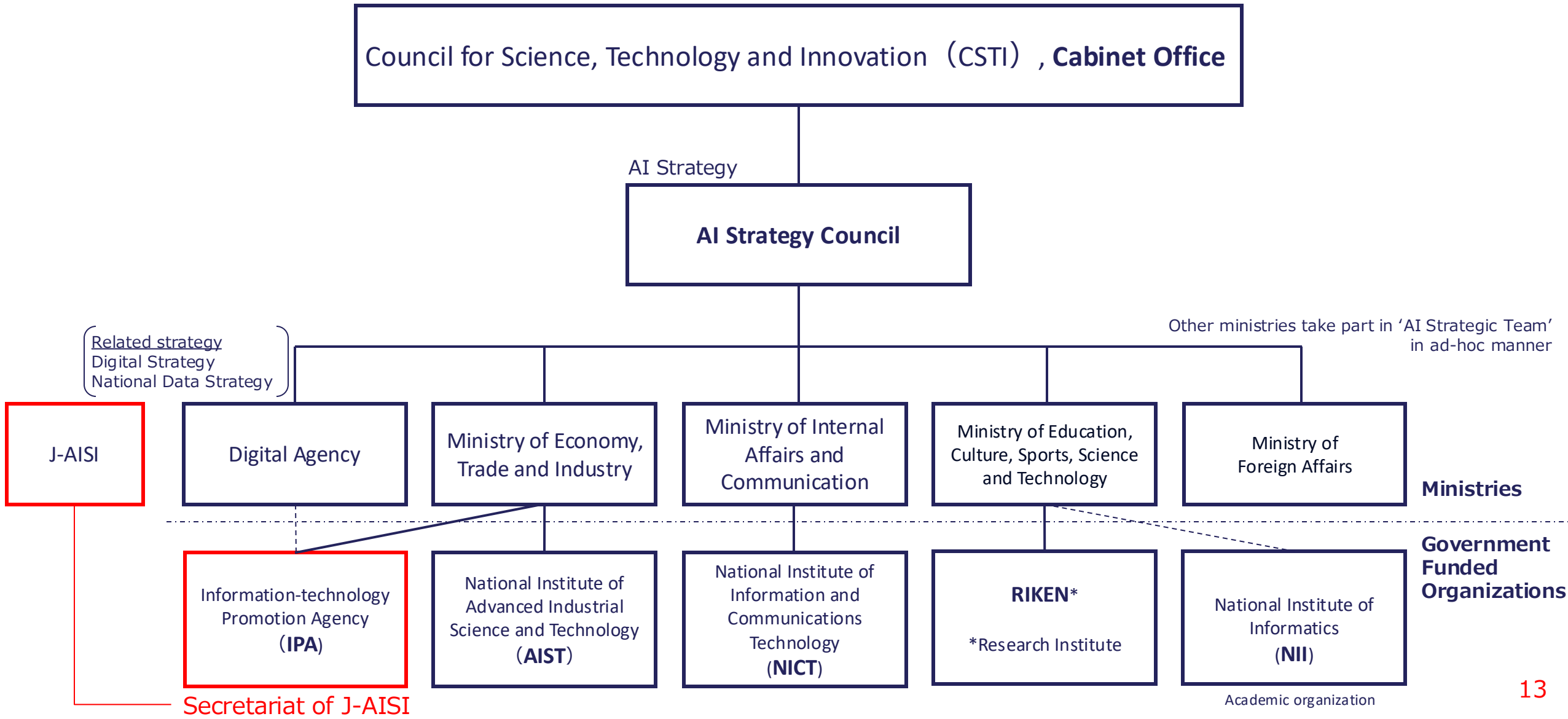
1. **Conducting the survey and study** for AI safety evaluation.
 - **Survey on standards of AI safety, checking tools, anti-disinformation technology, AI and Cybersecurity**
 - **Study on standards and guidance** regarding the AI safety
 - **Consideration of a testbed environment for AI** related to the above
2. **Considering the implementation methods** for AI safety evaluation.
3. **Collaborating Internationally** with related organizations in other countries.
(e.g. AI Safety Institute in the U.K. and the U.S.)

Immediate Activities and Deliverables

2024	International	J-AISI		Government
	EVENT	OUTPUT	EFFECT	
Apl		<ul style="list-style-type: none">• JP-U.S. Crosswalk1(4/30)		<ul style="list-style-type: none">• AI Guidelines for Businesses was published(4/19)
May	AI Safety Summit, Korea			
Jun	G7 Summit, Italy			
Jul		<ul style="list-style-type: none">• Japanese Translation of U.S. AI RMF(7/4)	 <ul style="list-style-type: none">✓ Provide basic information on AI safety✓ Enable global confirmation✓ Identify points of the evaluation✓ Understand the testing methodology	<ul style="list-style-type: none">• Integrated Innovation Strategy 2024 was published(6/4)
Aug				
Sep		<ul style="list-style-type: none">• Guide to Evaluation Perspectives(9/18)		
Oct		<ul style="list-style-type: none">• JP-U.S. Crosswalk2(9/18)		
Nov	International Network of AISIs Convening, USA	<ul style="list-style-type: none">• Guide to Red Teaming Methodology(9/25)		
⋮				

Our Team

AI Related Government Organization





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]



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

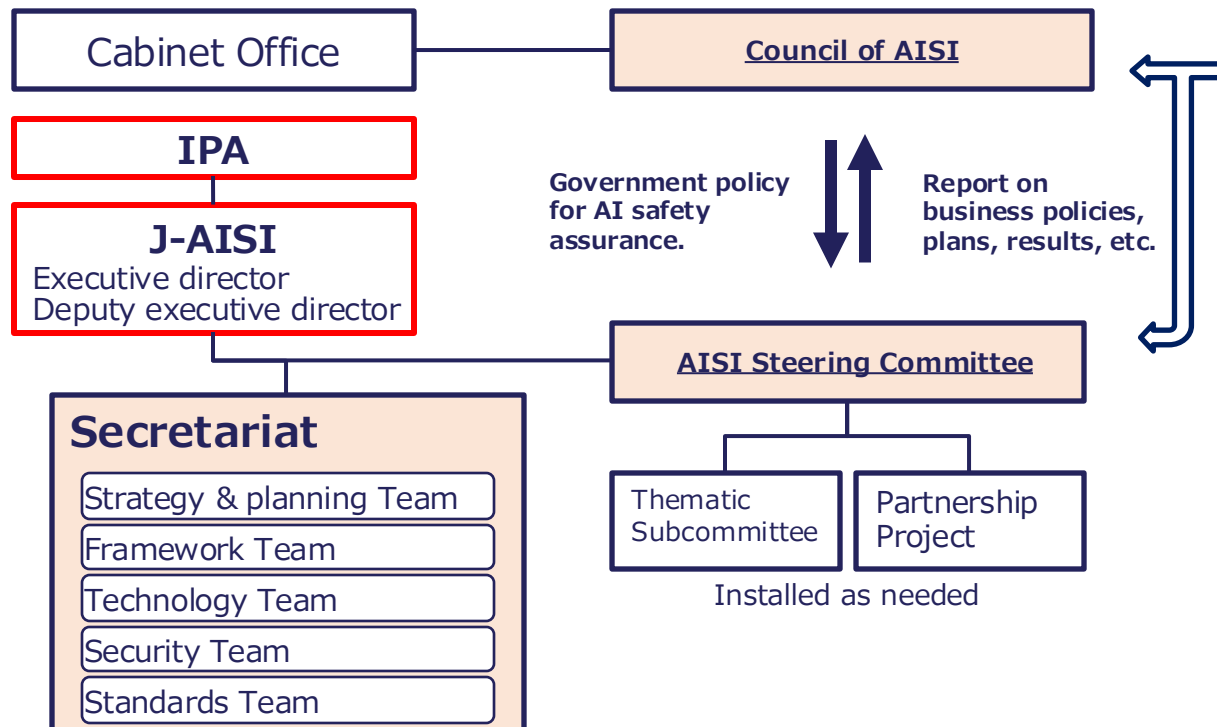
1999: Joined the Ministry of Posts and Telecommunications

2007: Ministry of Internal Affairs and Communications,
Telecommunications Bureau

2023: Cabinet Secretariat, Cabinet Cyber Security Center

2024: Deputy Director of J-AISI [Current]

- ♦ J-AISI is an organization formed with the cooperation of **10 relevant ministries** and **5 related organizations**.
 - **“Council”**, set up in Cabinet Office, deliberates on the important matters of J-AISI.
 - The **“AISAI Steering Committee”** within J-AISI reports to the Council.
 - **Secretariat**; mainly seconded from companies and ministries, and IPA staff.



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 Economy, Trade and Industry
- 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

♦ Strategy & Planning

- Create the strategies and plans, Manage budgets
- Understand the situation regarding AI safety
- Public relations (e.g., AISI web site)
- Recruitment and human resources development
- Coordination and support with related organizations

♦ Framework

- Support for the development of AI guidelines (e.g., **Crosswalk**)
- Support for international coordination of AI risk management frameworks
- Collection of information related to AI governance and Technical Advice
- Support for consideration of the nature of **certification and accreditation**

◆ Technology

- Organizing **Red Team** Implementation Methodology
- Collect information and provide advice that contributes to the development of technology-related standards, guidelines, etc.
 - **Synthetic content, disinformation and misinformation**
 - Bias, data checking, history management
- Consideration of tools such as test beds, etc.

◆ Security

- Consideration of security measures for AI
- AI-based support for countermeasures against security events

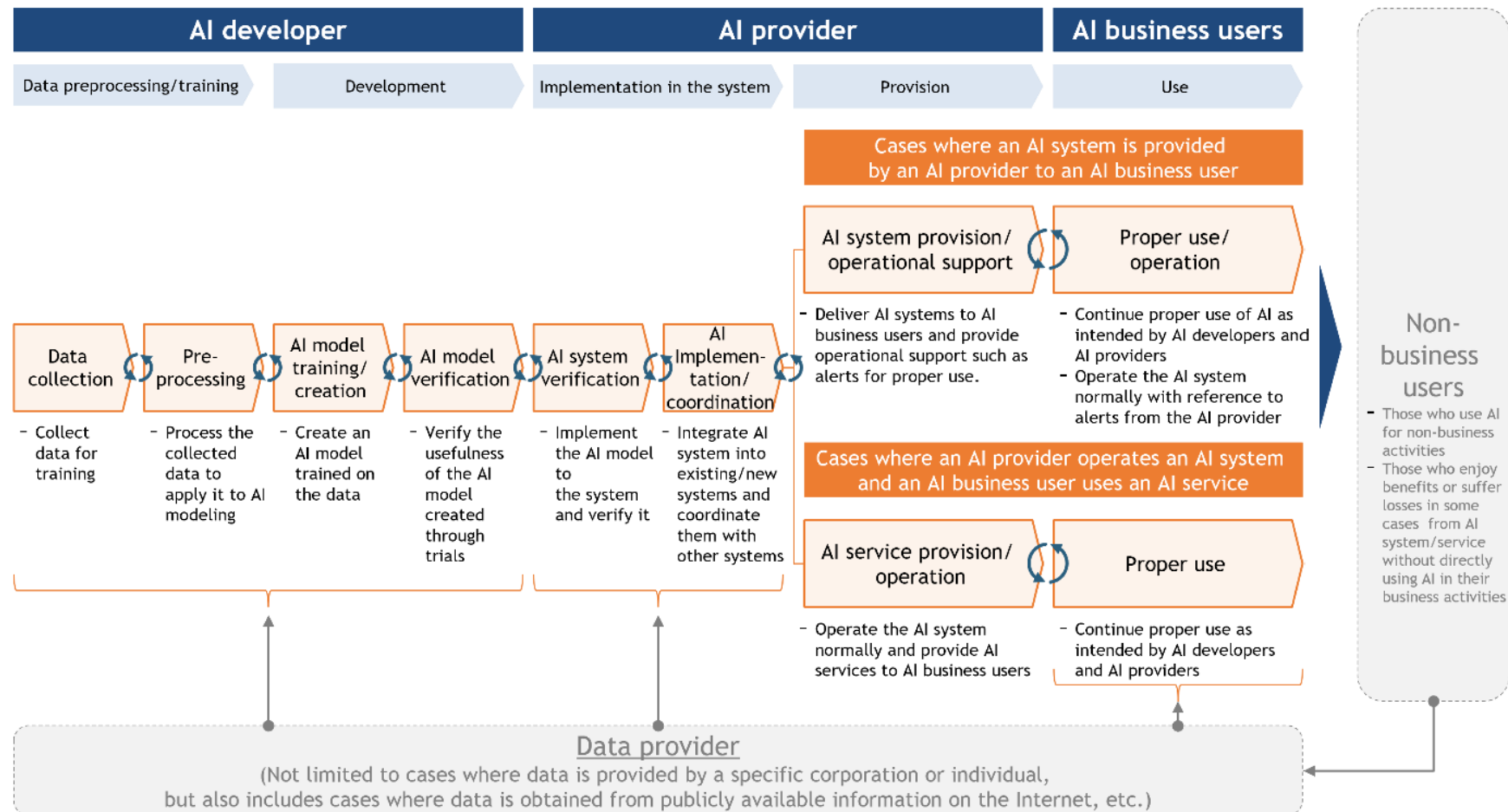
◆ Standards

- Support for the promotion of ISO SC42
- Collection of other standard information

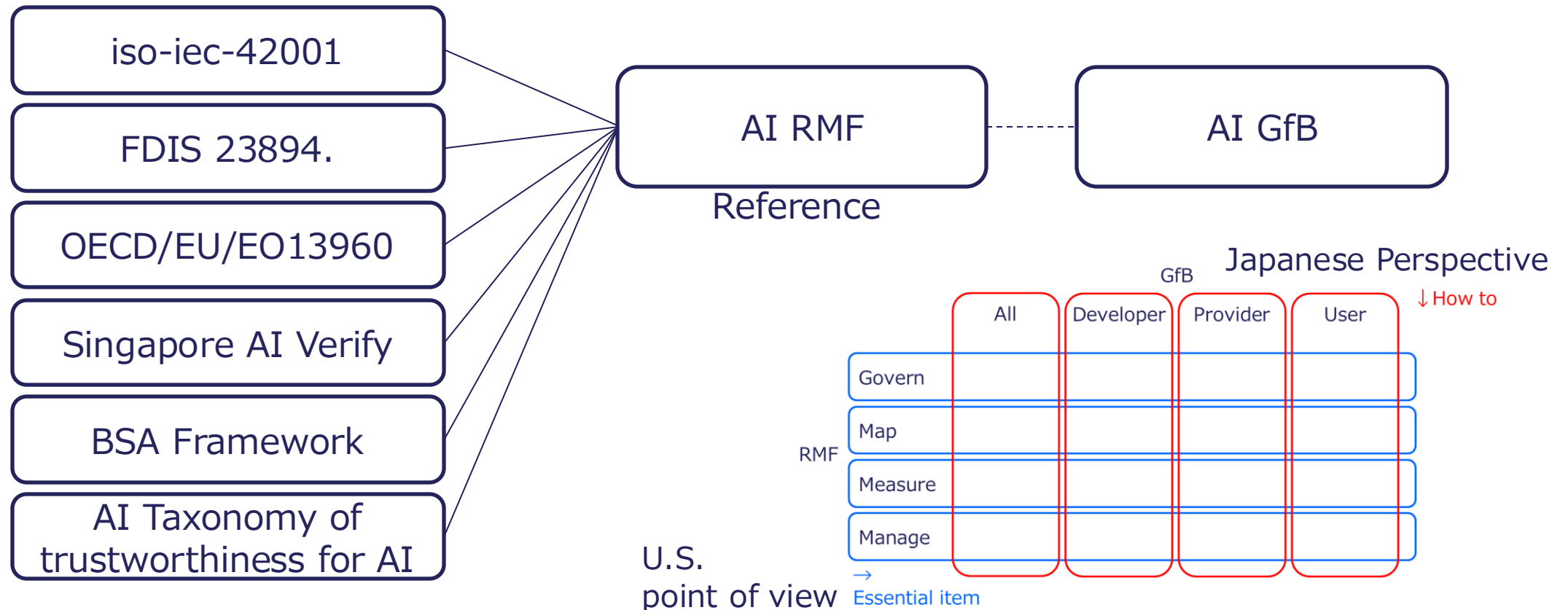
Recent Developments related to AISJ Activities

AI Guidelines for Business

- ◆ Clarify what each stakeholder should address in the flow of AI utilization



- ◆ Confirmation of the interrelationship between the U.S. NIST AI Risk Management Framework (RMF) and the Japanese AI Guidelines for Business (GfB)



Overview of Japan-U.S. Crosswalk

◆ Crosswalk 1 (Released on April 30th)

[Crosswalk1.pdf](#)

◆ Crosswalk 2 (Released on September 18th)

[Crosswalk2.pdf](#)



As a first step of JPN-US crosswalk, J-AISI and NIST together publish Crosswalk 1-Terminology. We look forward to advancing the crosswalk, aiming at promoting interoperability of JPN-US AI governance frameworks.

aisi.go.jp/international/

18:30 · 2024/04/30 · 11K Views

Crosswalk 1 – Terminology NIST AI Risk Management Framework (NIST AI RMF) and Japan AI Guidelines for Business (AI GfB)	
NIST AI RMF 1.0 - Characteristics of Trustworthy AI Systems	Japan AI GfB - Common Guiding Principles
Valid & Reliable – (Includes accuracy and robustness) Validation: “confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled” ¹ Reliability: “ability of an item to perform as required, without failure, for a given time interval, under given conditions” ² Accuracy: “closeness of results of observations, computations, or estimates to the true values or the values accepted as being true” ² Robustness: “ability of a system to maintain its level of performance under a variety of circumstances” ²	Validation: (There is no definition for validation. Instead, as an element of transparency, the AI GfB indicates the importance of ensuring the verifiability of the AI systems and services as necessary and technically possible.) Reliability: The AI works satisfactorily for the requirements, including the accuracy of its output Accuracy: The AI works satisfactorily for the requirements Robustness: Maintaining performance levels under a variety of conditions and avoiding significantly incorrect decisions regarding unrelated events <u>AI GfB Context</u> 2) Safety (Includes accuracy, reliability, and robustness) (1) Consideration for human life, body, property and mind as well as

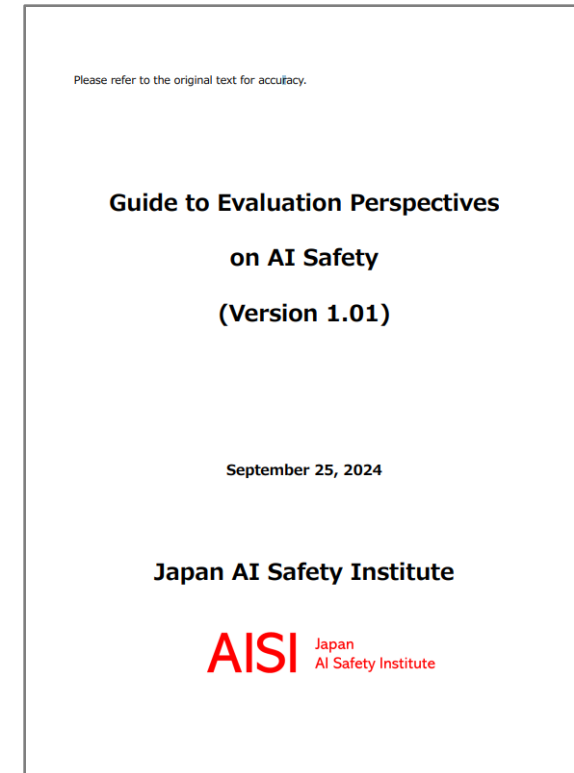
Crosswalk 2 – Concepts NIST AI Risk Management Framework and Japan AI Guidelines for Business

The NIST AI Risk Management Framework 1.0 (NIST AI RMF)¹ is a voluntary resource created to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. The Japan AI Guidelines for Business (Japan AI GfB)² is a framework that promotes innovation and the reduction of risks across the lifecycle by encouraging AI business actors to fully recognize AI risks based on international trends and stakeholders’ concerns. The purpose of this crosswalk is to compare and contrast the key concepts addressed in the NIST AI RMF with the concepts addressed in the Japan AI GfB.

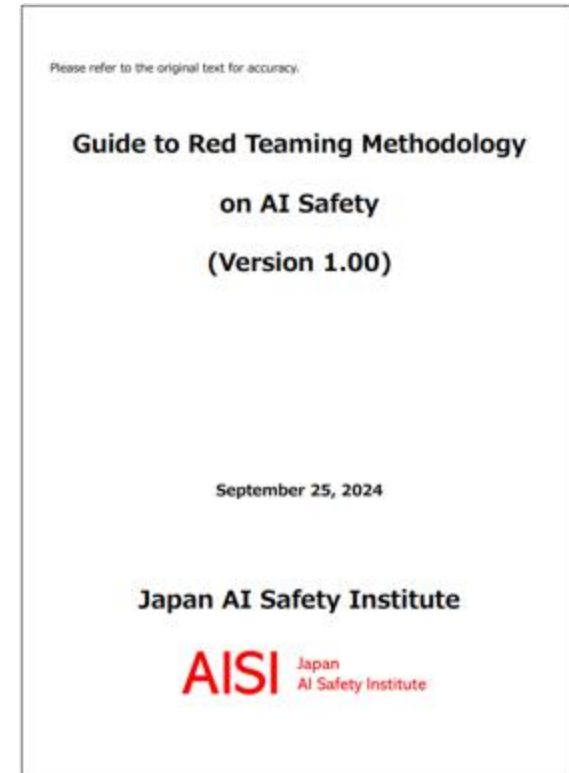
Table 1 below presents the crosswalk mapping and analysis. The following list describes each column in the table.

Topic	Conceptual topics derived from the NIST AI RMF Playbook ³ used to anchor the comparative analysis of this crosswalk
NIST AI RMF References	NIST AI RMF subcategories linked to each topic by the Playbook
Japan AI GfB References	Japan AI GfB sections (including appendices ⁴) identified by Japan as addressing the topic
Notable Similarities & Differences	Observations on how the NIST AI RMF and the Japan AI GfB align or differ on the treatment of each topic ⁵

- ◆ The significance and use of the Evaluation Perspectives Guide on AI Safety.
 - Provides a basic approach to evaluating the safety of AI Systems, and it is intended to be used as a reference for operators when developing and providing AI.
 - Including:
 - Risks and evaluation items assumed in the safety assessment.
 - Conductors and timing of the evaluation.
 - Overview of Evaluation Methodology.
- ◆ This guide is the first step toward realizing safe, secure, and reliable AI.
 - We hope that it will contribute to maintaining and improving safety in the development and provision of AI in the future.



- ◆ The significance and use of the Red Teaming Methodology Guide on AI Safety.
 - Provides a basic considerations for the red teaming method (one of the methods used to evaluate the safety of AI systems), and it is intended to be used as a reference for operators when developing and providing AI.
 - 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 the first step toward realizing safe, secure, and reliable AI. We hope that it will contribute to maintaining and improving safety in the development and provision of AI in the future.



♦ AISI-related collaborations

- Stanford University AI Symposium (Stanford, April 16)
 - 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)
 - 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.
- Asia Tech x Summit in Singapore (online, May 31)
 - Panel discussion with the director of U.S. AISI.
- UN Future Summit (UN, September 22)
- UN Global Compact Leaders Summit 2024 (UN, September 24)
- AISI International Network Convening (San Francisco, November 10-11)

♦ Active exchange of views with AI-related businesses operators and organizations

- Discussions with the U.S., U.K., EU, Singapore, Australia, and South Korea.
- Discussions with executives of business operators.
- Participation in GPAI workshop (Paris, May 22-23)



AI Global Forum,
held in conjunction with AI Seoul Summit



United Nations
Summit of the Future

- ◆ Cooperation with the private sector
 - Consideration of Establishment of “Business Demonstration Working Group”.
 - Announced at the 3rd meeting of Japan AISI Steering Committee.
- ◆ Expansion of Research and Studies
 - Research for the revision of the Evaluation Perspectives Guide and the Red Teaming Methodology Guide.
 - Expanding the scope to include multimodal infrastructure models.
 - Research for the Business Demonstration WG initiatives.
 - To establish part of the evaluation environment for AI safety ahead of others and develop the environment for WG activities.
 - Examination on automatic evaluation of AI safety
 - Consideration of automation/laborsaving of assessments to make AI safety assessments widely available to the public.
- ◆ Major Events
 - 2024 December 11(WEB) [How should Japanese companies prepare for the EU-AI Act?](#)
 - 2025 February 10-11(Paris) [Artificial Intelligence Action Summit](#)

AISI

Japan AI Safety Institute