Overview of the AI Safety Institute (AISI)

2024-11-01

AI Safety Institute (AISI)

AISI Japan AI Safety Institute

Background of AISI

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 AISI
- **February 14, 2024**
  - AI Safety Institute (AISI) was established

About AISI

AISI Japan AI Safety Institute

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

Role and Scope of AISI

AISI Japan AI Safety Institute

- **Role**
  - **AISI supports the government** by conducting surveys on AI safety, examining evaluation methods, and creating standards.
  - **As a hub for AI safety in Japan**, AISI will consolidate the latest information in industry and academia, and promote collaboration among related companies and organizations.
  - **Collaborate with AI safety-related organizations**.
    - AISI is not an R&D organization.
- **Scope**
  - **Set the scope flexibly** in the following AI related issues, while considering **global trends**.
    - **Social Impact**
    - **governance**
    - **AI System**
    - **contents**
    - **data**

Business

AISI Japan AI Safety Institute

1) **Consideration of surveys and standards** for AI safety assessment

   (i) **Surveys on standards of AI safety, checking tools, anti-disinformation technology, AI and cybersecurity**

   (ii) **Consideration of standards and guidance** related to AI safety

   (iii) **Consideration of a testbed environment for AI** related to the above

2) **Consideration of implementation methods** for AI safety assessment

3) **International collaboration** with related organizations in other countries (such as the AI Safety Institute in the U.K. and the U.S.)

Immediate Activities and Deliverables

AISI Japan AI Safety Institute

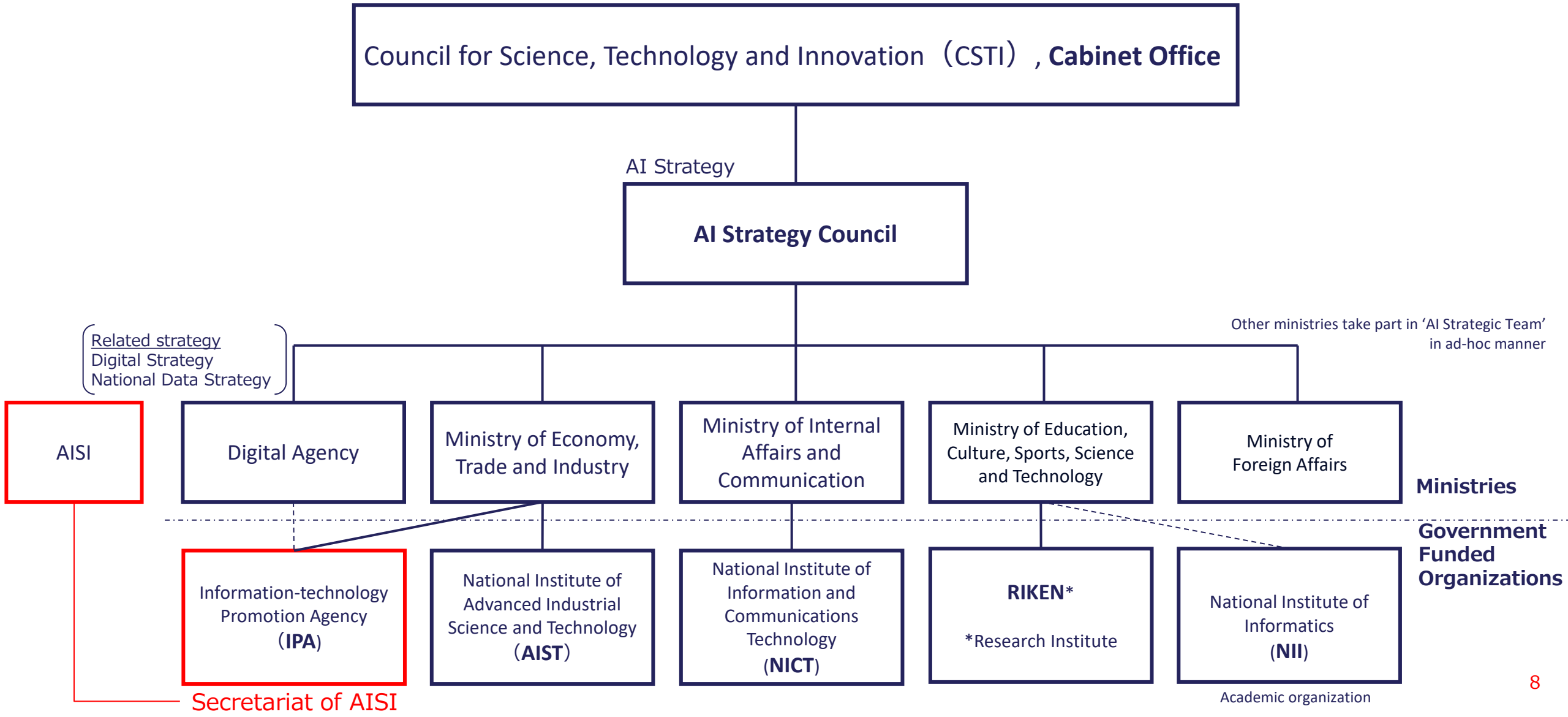
	International	AISI		Government
2024	EVENT	OUTPUT	EFFECT	
Apl		• JP-US Crosswalk1(4/30)		• AI Guidelines for Businesses was published(4/19)
May	AI Safety Summit, Korea		✓ Provide basic information on AI safety ✓ Enable global confirmation ✓ Identify points of the evaluation ✓ Understand the testing methodology	
Jun	G7 Summit, Italy			
Jul		• Japanese Translation of U.S. AI RMF(7/4)		• Japanese Government strategies were published
Aug				
Sep		• Guide to Evaluation perspectives(9/18) • JP-US Crosswalk2(9/18) • Guide to Red Teaming Methodology(9/25)		
Oct	International Network of AISIs Convening, USA			
⋮				

AISI Japan AI Safety Institute

Our team

AI Related Government organization

AISI Japan AI Safety Institute



Executive Team

Executive Director

Akiko Murakami

Deputy Executive Director
Secretary General

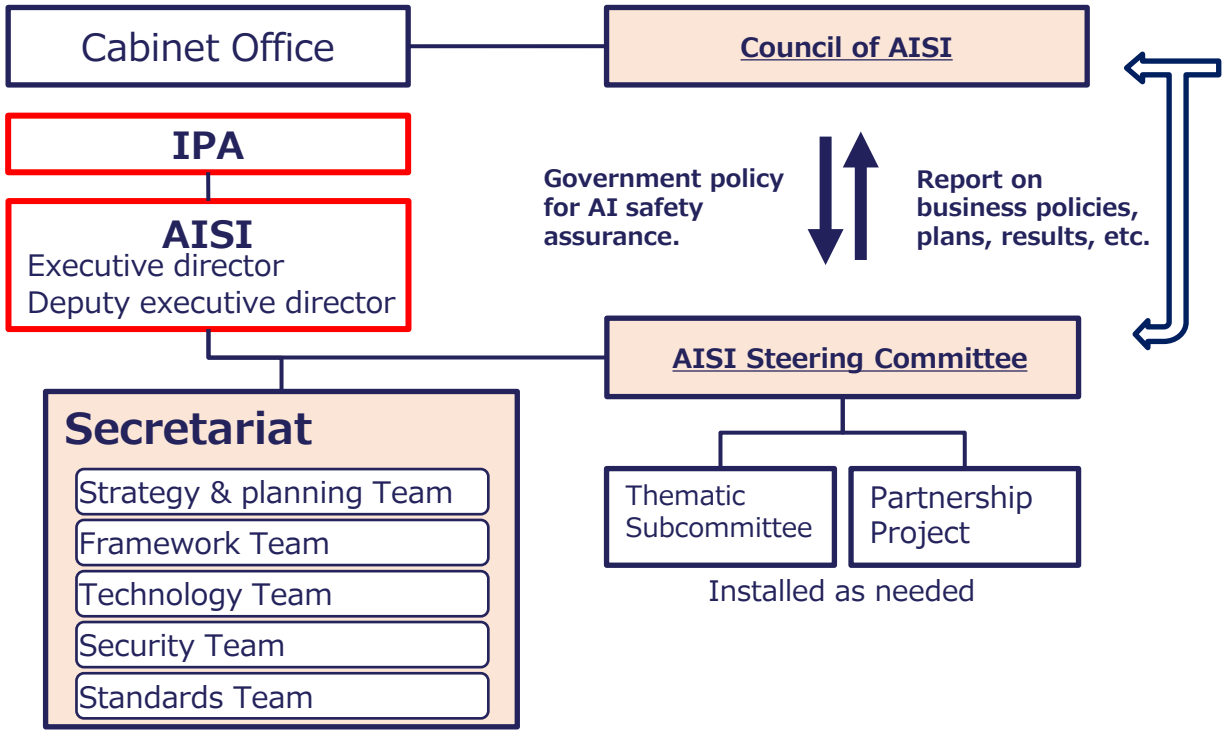
Kenji Hiramoto

Deputy Executive Director

Hideyuki Teraoka

AISI Structures

- 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 AISI.
  - The **"AISI Steering Committee"** within 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

Items to be implemented at AISI

AISI Japan AI Safety Institute

- **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**

Items to be implemented at AISI

AISI Japan AI Safety Institute

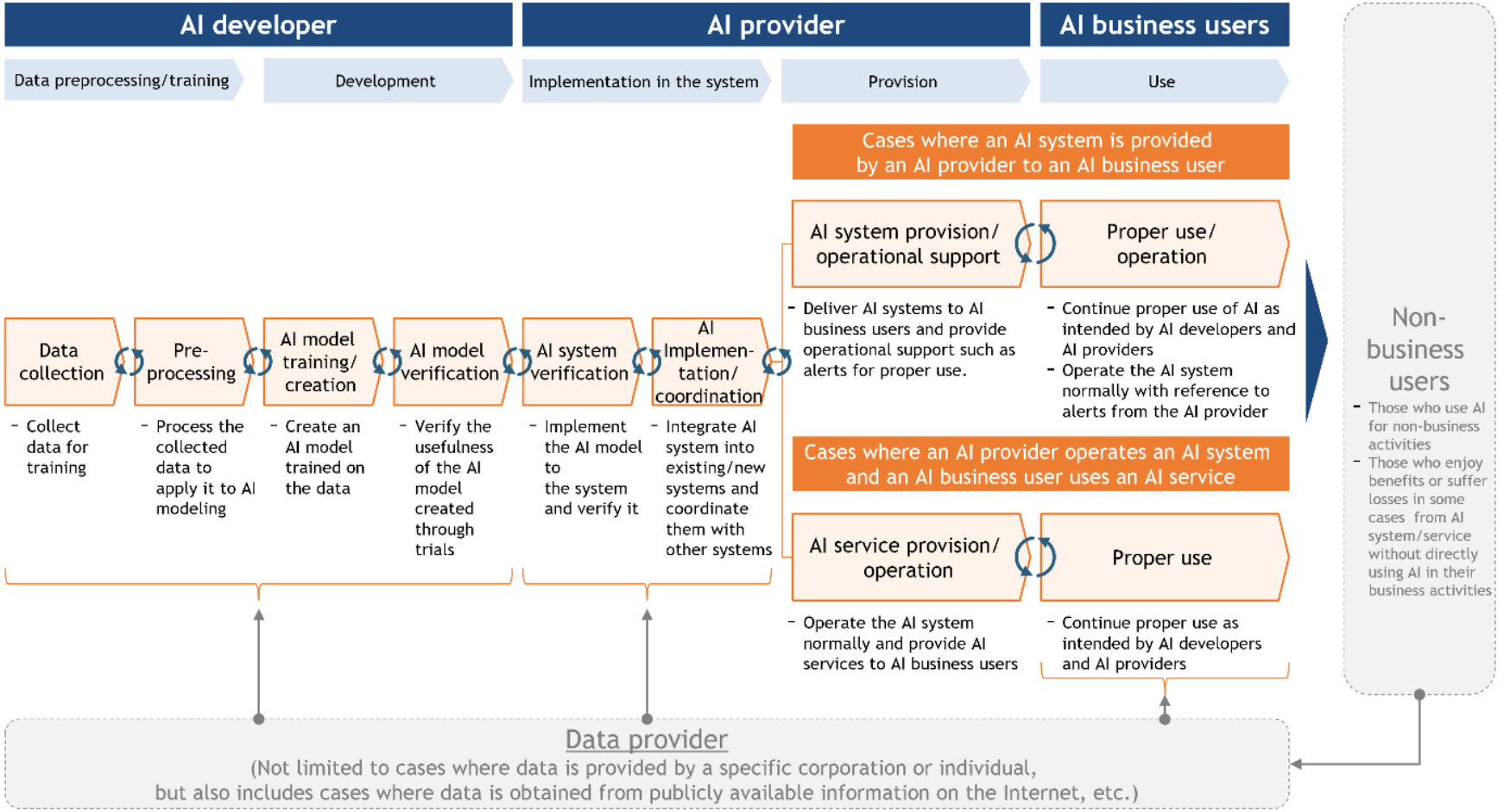
- **Technology**
  - Organizing **Red Team** Implementation Methodology
  - Collect information and provide advice that contributes to the development of technology-related standards, guidelines, etc.
    - **Synthetic content, disinformation and misinformation**
    - Bias, data checking, history management
  - 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 AISI activities

AI Guidelines for Business

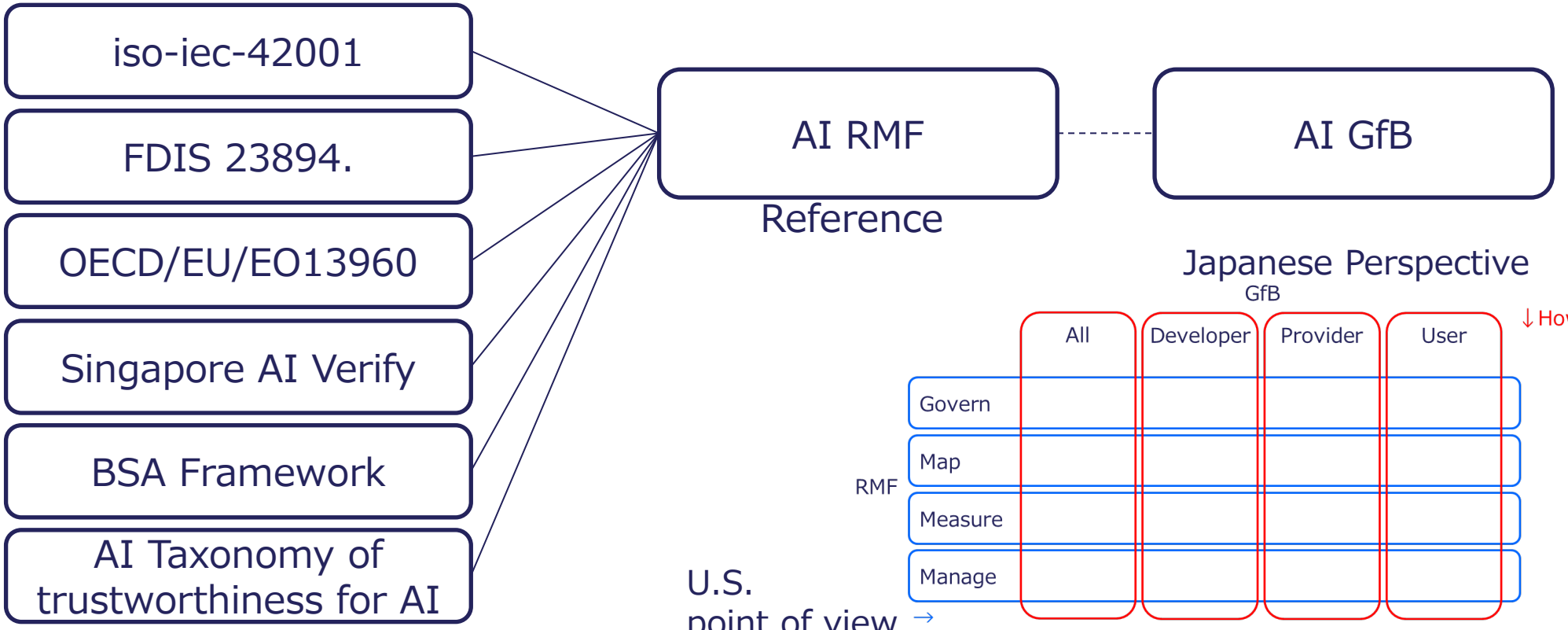
AISI Japan AI Safety Institute

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



Japan-U.S. Crosswalk

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

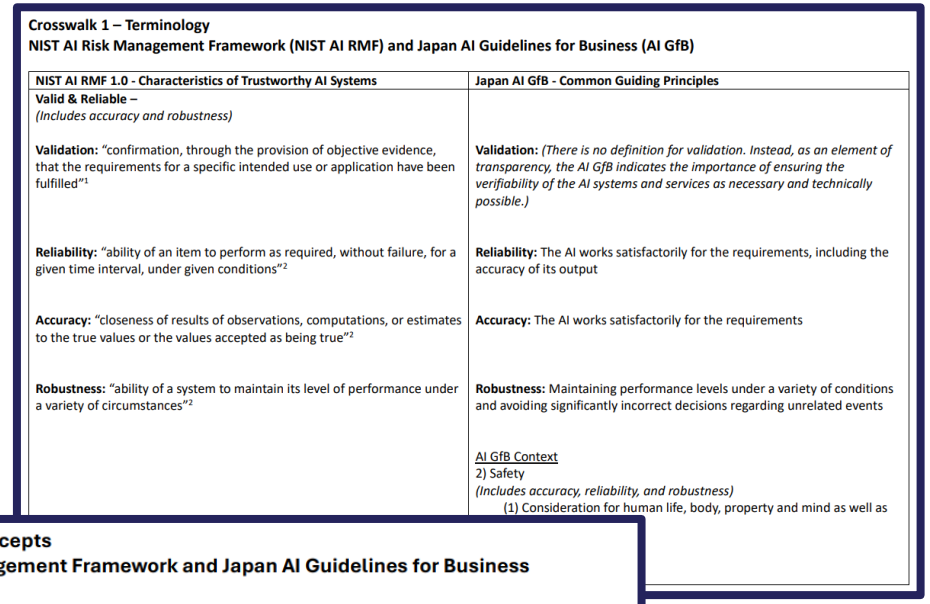


Overview of Japan-U.S. Crosswalk

AISI Japan AI Safety Institute

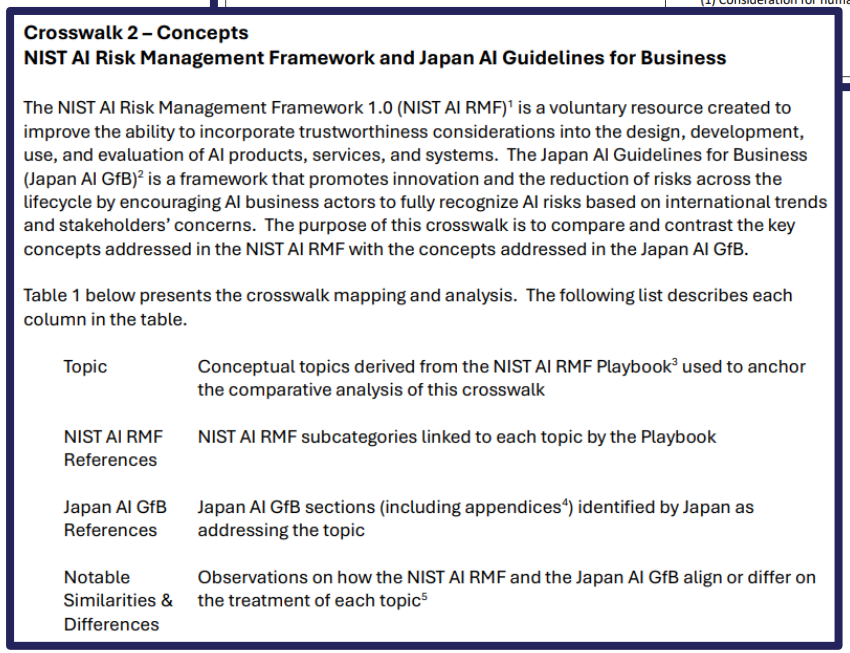
- Crosswalk 1 (Released on April 30th)
  https://aisi.go.jp/assets/pdf/AISI_Crosswalk1_RMF_GfB_ver1.0.pdf

- Crosswalk 2 (Released on September 18th)
  https://aisi.go.jp/assets/pdf/AISI_Crosswalk2_RMF_GfB_ver1.0.pdf



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]	**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: "ability of an item to perform as required, without failure, for a given time interval, under given conditions"[2]	**Reliability:** The AI works satisfactorily for the requirements, including the accuracy of its output
Accuracy: "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true"[2]	**Accuracy:** The AI works satisfactorily for the requirements
Robustness: "ability of a system to maintain its level of performance under a variety of circumstances"[2]	**Robustness:** Maintaining performance levels under a variety of conditions and avoiding significantly incorrect decisions regarding unrelated events
	AI GfB Context 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.

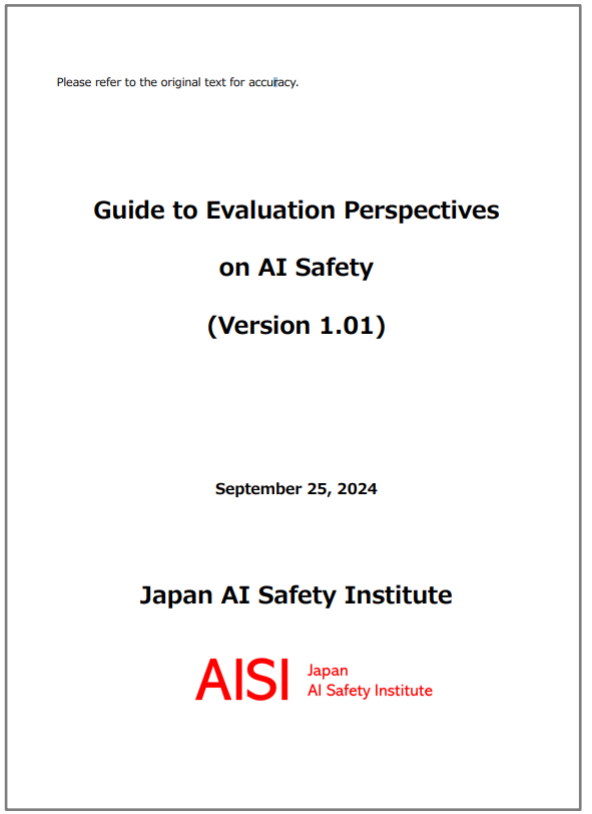
Table 1 below presents the crosswalk mapping and analysis. The following list describes each column in the table.

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]

Guide to Evaluation Perspectives on AI Safety

AISI Japan AI Safety Institute

■ The significance and use of the Evaluation Perspectives Guide on AI Safety.


Please refer to the original text for accuracy.

Guide to Evaluation Perspectives on AI Safety

(Version 1.01)

September 25, 2024

Japan AI Safety Institute

AISI Japan AI Safety Institute

➢ 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.

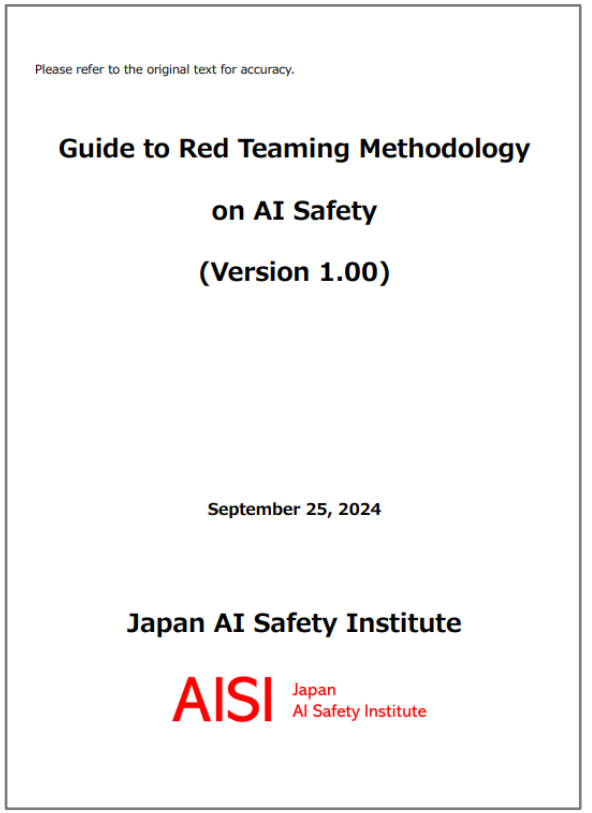
➢ This guide is the first step toward realizing safe, secure, and reliable AI.
We hope that it will contribute to maintaining and improving safety in the development and provision of AI in the future.

https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.01_en.pdf

Red Teaming Methodology Guide

AISI Japan AI Safety Institute

■ The significance and use of the Red Teaming Methodology Guide on AI Safety.


Please refer to the original text for accuracy.

Guide to Red Teaming Methodology

on AI Safety

(Version 1.00)

September 25, 2024

Japan AI Safety Institute

AISI Japan AI Safety Institute

- 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.

https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf

International Cooperation

AISI Japan AI Safety Institute

◆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)
  - Discussions with AI leaders of various countries on AI safety.

AI Global Forum,
held in conjunction with AI Seoul Summit

◆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)

United Nations
Summit of the Future

Plans for future initiatives

AISI Japan AI Safety Institute

- 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 in which AISI participates
  - November 8 ISS square Workshop
  - November 11 Towards Building Multicultural and Multilingual Safe Large Language Models
  - November 20-21 International Network of AISIs Convening

AISI

Japan AI Safety Institute