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

2025-01-06

AISI Japan
AI Safety Institute

AI Strategy

Integrated Innovation Strategy 2024

- 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

Integrated Innovation Strategy 2024.pdf

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)

Background of J-AISI

- 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

Initiatives to ensure AI safety in each country

- **United States**
  - AISI established within the National Institute of Standards and Technology (NIST).
  - Fundamentally led by the private sector, with strong promotion of collaboration with the consortium of private companies (AISIC).
  - Staffs are approximately 30 members. Aiming for around 80 members.
- **United Kingdom**
  - AISI established within the Department of Science, Innovation and Technology (DSIT).
  - The government is leading the promotion of the evaluation and testing of AI safety.
  - The scale is about 100 people and planning to increase the engineers. Recently opened San Francisco branch.
- **EU**
  - Promoting the utilization and safety of AI at the AI Office of the European Commission (EC). Also responsible for developing and promoting AI law.
  - Approximately 60 people.
- **Singapore**
  - The Digital Trust Center at Nanyang Technological University (NTU) has designated AISI in Singapore.
  - Provides safety evaluation test tools, etc., with the aim of international standardization of large-scale language models (LLMs).
- **Canada**
  - AISI established with the cooperation of domestic organizations.
- **Republic of Korea**
  - Established in November 2024.
  - Aims to become a hub of Asia.
- **Australia**
  - A national research institute is charge of the AISI function.

About J-AISI

- **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 and Scope of J-AISI

- ## 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.
  - In addition, an **international consensus will be established** by **collaborating with AI safety-related organizations** in other countries.
- ## Scope
  - **Set the scope flexibly** in the following AI related issues, while considering **global trends**.
    - **Social Impact**
    - **Governance**
    - **AI System**
    - **Contents**
    - **Data**

Actions

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

AISI Japan AI Safety Institute

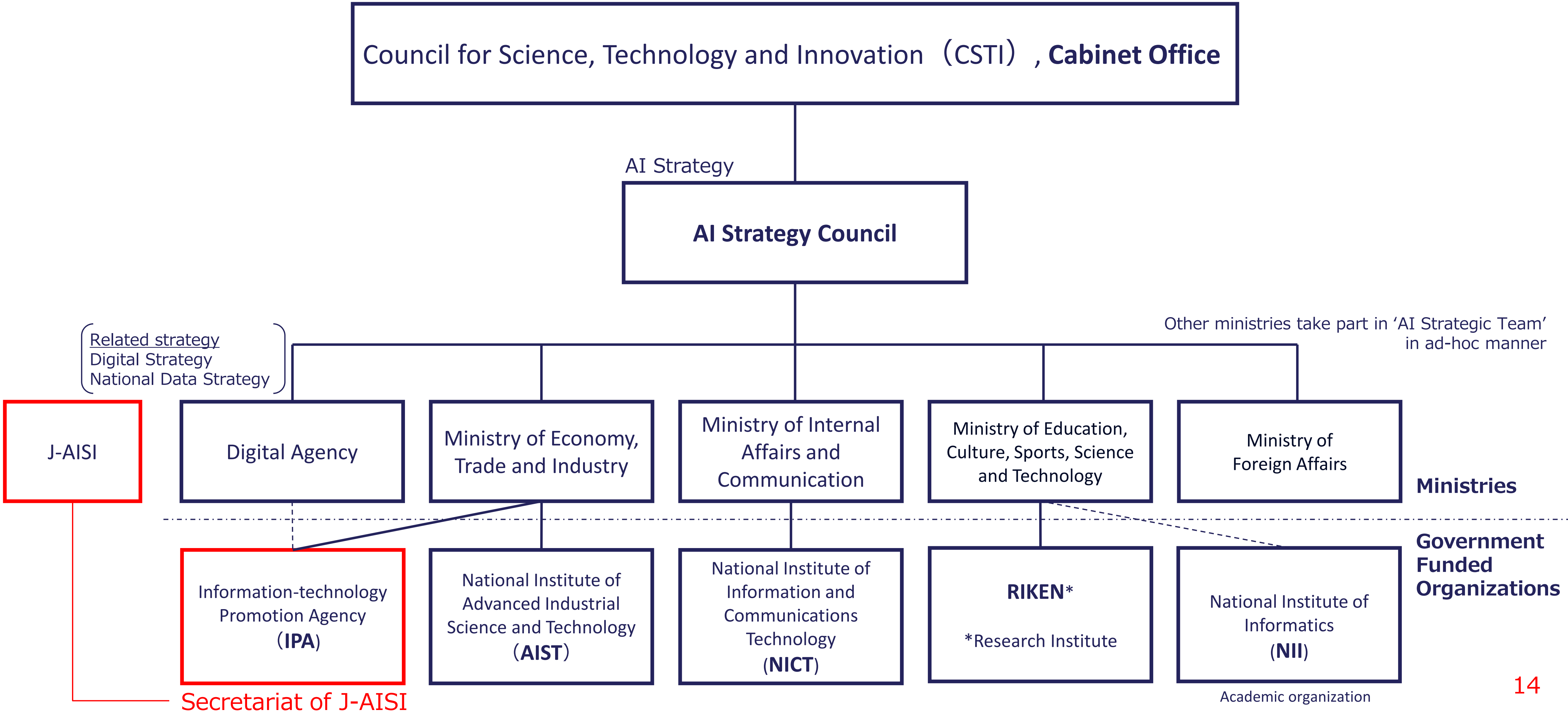
2024	International	J-AISI		Government
	EVENT	OUTPUT	EFFECT	
Apl		• **JP-U.S. Crosswalk1**(4/30)	✓ Provide basic information on AI safety ✓ Enable global confirmation ✓ Identify points of the evaluation ✓ Understand the testing methodology	• **AI Guidelines for Businesses** was published(4/19)
Mar	AI Safety Summit, Korea			
Jun	G7 Summit, Italy			• **Integrated Innovation Strategy 2024** was published(6/4)
Sep				
Aug		• Japanese Translation of U.S. AI RMF(7/4)		
Sep		• **Guide to Evaluation Perspectives**(9/18) • **JP-U.S. Crosswalk2**(9/18)		
Oct		• **Guide to Red Teaming Methodology**(9/25)		
Nov	International Network of AISIs Convening, USA			
Dec				

Upcoming Schedule

Jan : Publication of Annual Report
Feb : AI Action Summit, France
Mar : Update of guide by AISI, summary of survey projects, update of AI Guidelines for Businesses guidelines

Our Team

AI Related Government Organization



Council for Science, Technology and Innovation（CSTI）, **Cabinet Office**

AI Strategy

AI Strategy Council

Related strategy
Digital Strategy
National Data Strategy

Other ministries take part in ‘AI Strategic Team’ in ad-hoc manner

J-AISI

Digital Agency

Ministry of Economy, Trade and Industry

Ministry of Internal Affairs and Communication

Ministry of Education, Culture, Sports, Science and Technology

Ministry of Foreign Affairs

Ministries

Information-technology Promotion Agency（**IPA**）

National Institute of Advanced Industrial Science and Technology（**AIST**）

National Institute of Information and Communications Technology（**NICT**）

RIKEN*

*Research Institute

National Institute of Informatics（**NII**）

Academic organization

Government Funded Organizations

Secretariat of J-AISI

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]

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]

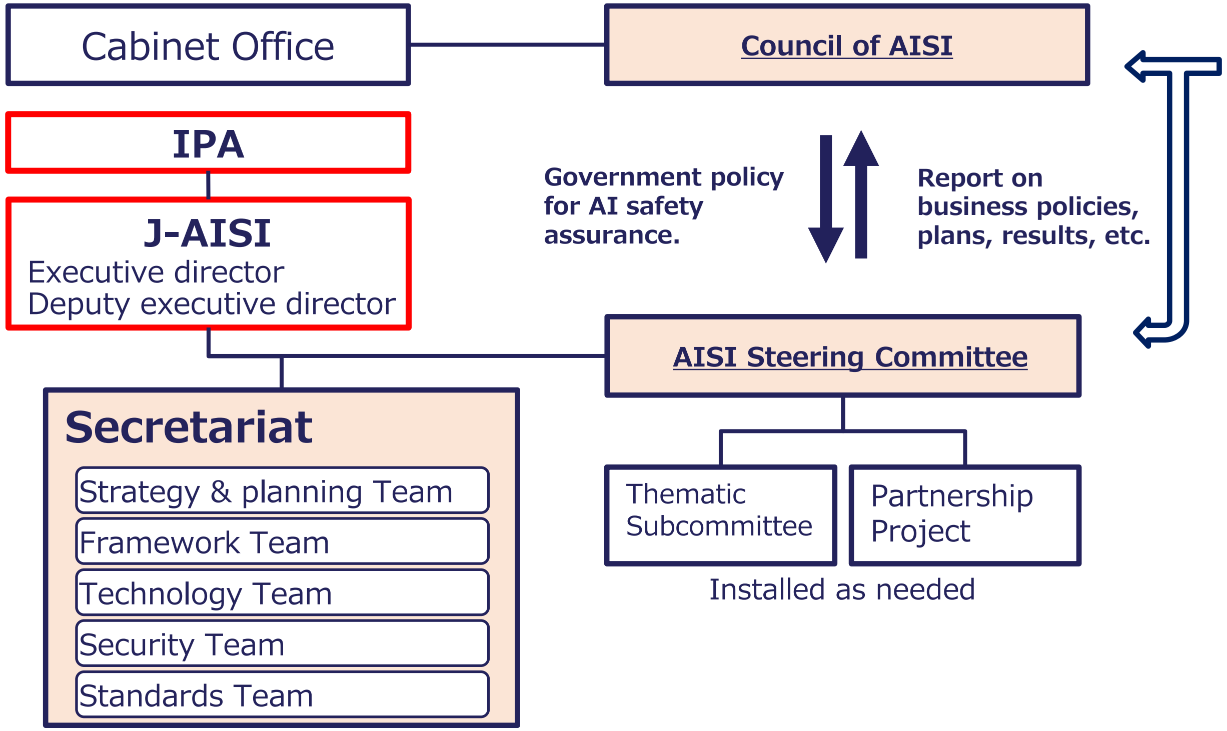
Deputy Executive Director

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 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 J-AISI.
  - The **“AISI 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

Items to be implemented at J-AISI

- **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**

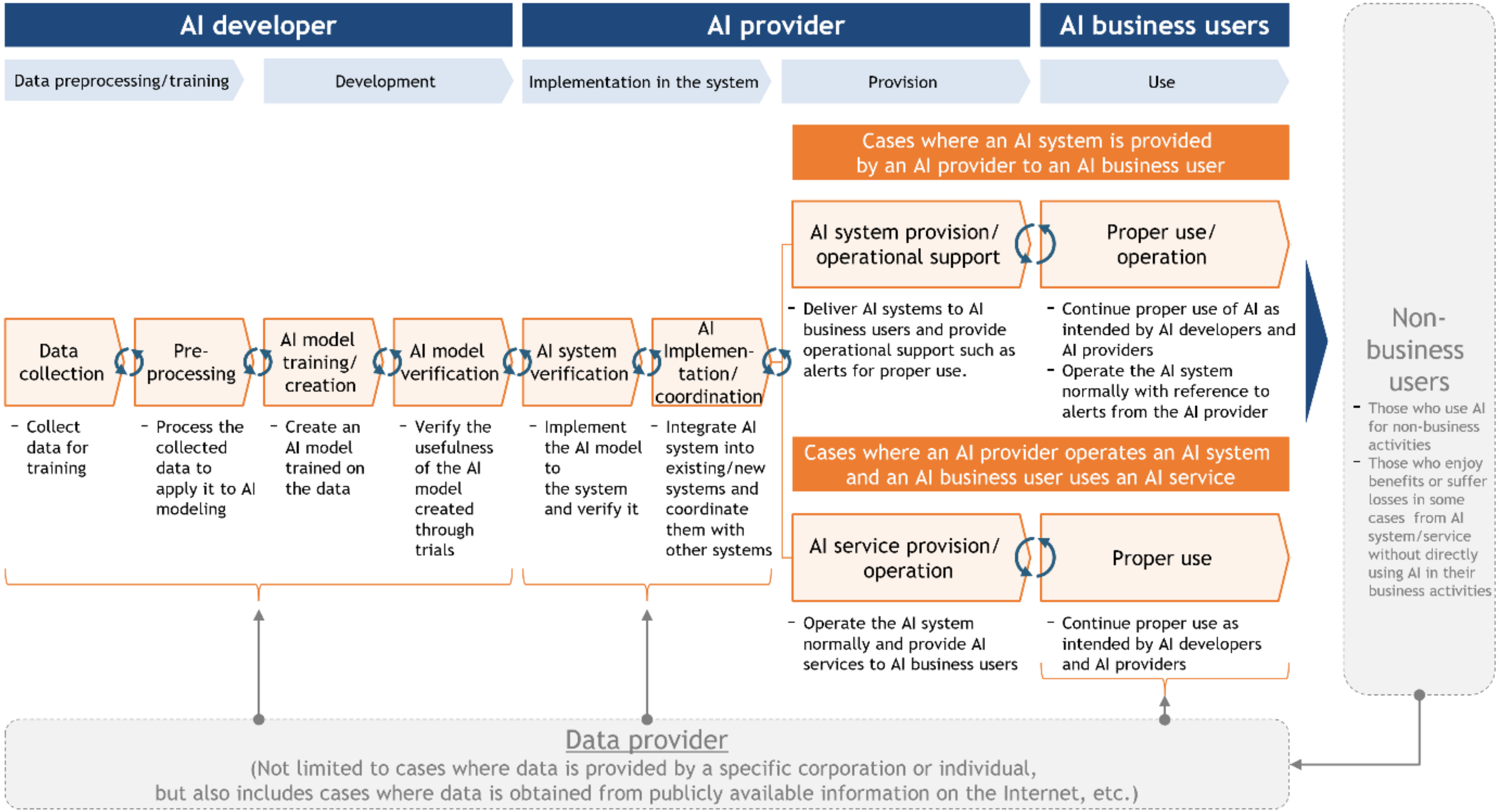
Items to be implemented at J-AISI

- **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

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



AI developer

AI provider

AI business users

Data preprocessing/training

Development

Implementation in the system

Provision

Use

Data collection
- Collect data for training

Pre-processing
- Process the collected data to apply it to AI modeling

AI model training/ creation
- Create an AI model trained on the data

AI model verification
- Verify the usefulness of the AI model created through trials

AI system verification
- Implement the AI model to the system and verify it

AI Implemen-tation/ coordination
- Integrate AI system into existing/new systems and coordinate them with other systems

Cases where an AI system is provided by an AI provider to an AI business user

AI system provision/ operational support
- Deliver AI systems to AI business users and provide operational support such as alerts for proper use.

Proper use/ operation
- 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

Cases where an AI provider operates an AI system and an AI business user uses an AI service

AI service provision/ operation
- Operate the AI system normally and provide AI services to AI business users

Proper use
- Continue proper use as intended by AI developers and AI providers

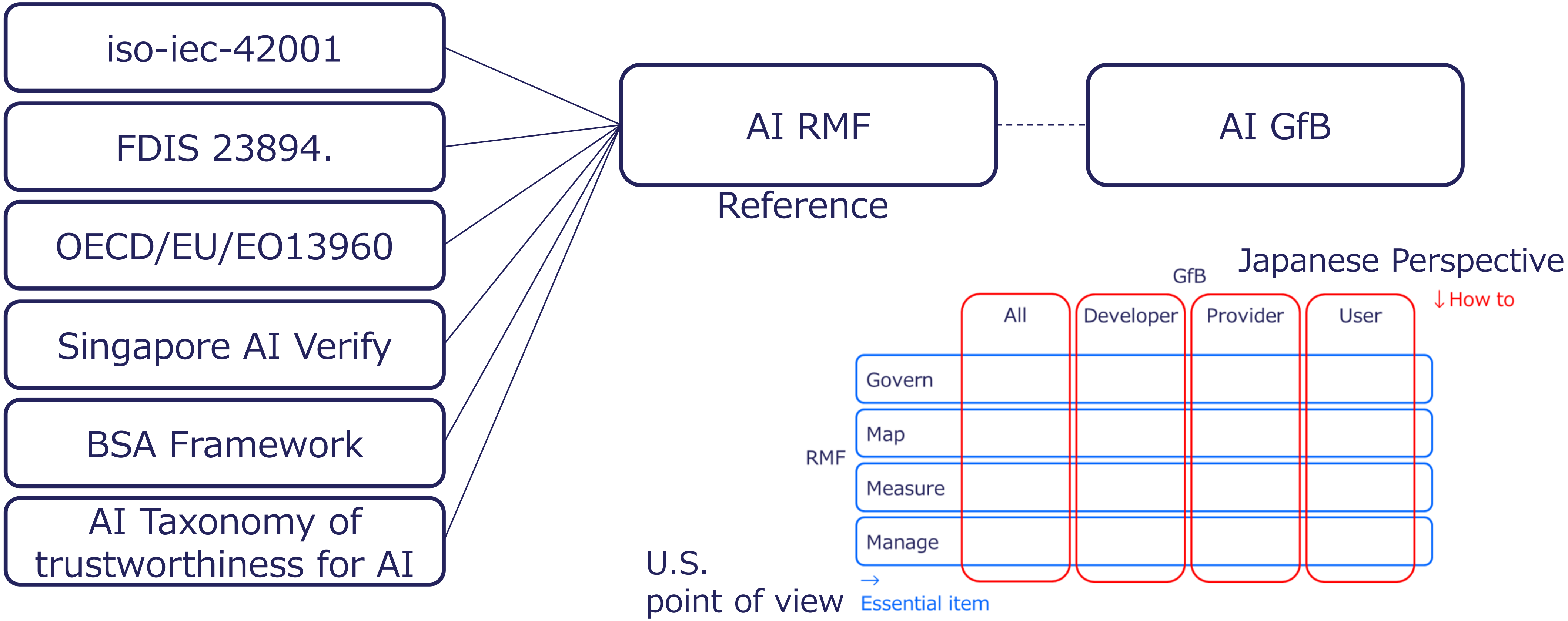
Non-business users
- Those who use AI for non-business activities
- Those who enjoy benefits or suffer losses in some cases from AI system/service without directly using AI in their business activities

Data provider

(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.)

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)



iso-iec-42001

FDIS 23894.

OECD/EU/EO13960

Singapore AI Verify

BSA Framework

AI Taxonomy of trustworthiness for AI

AI RMF

Reference

AI GfB

GfB

Japanese Perspective

↓How to

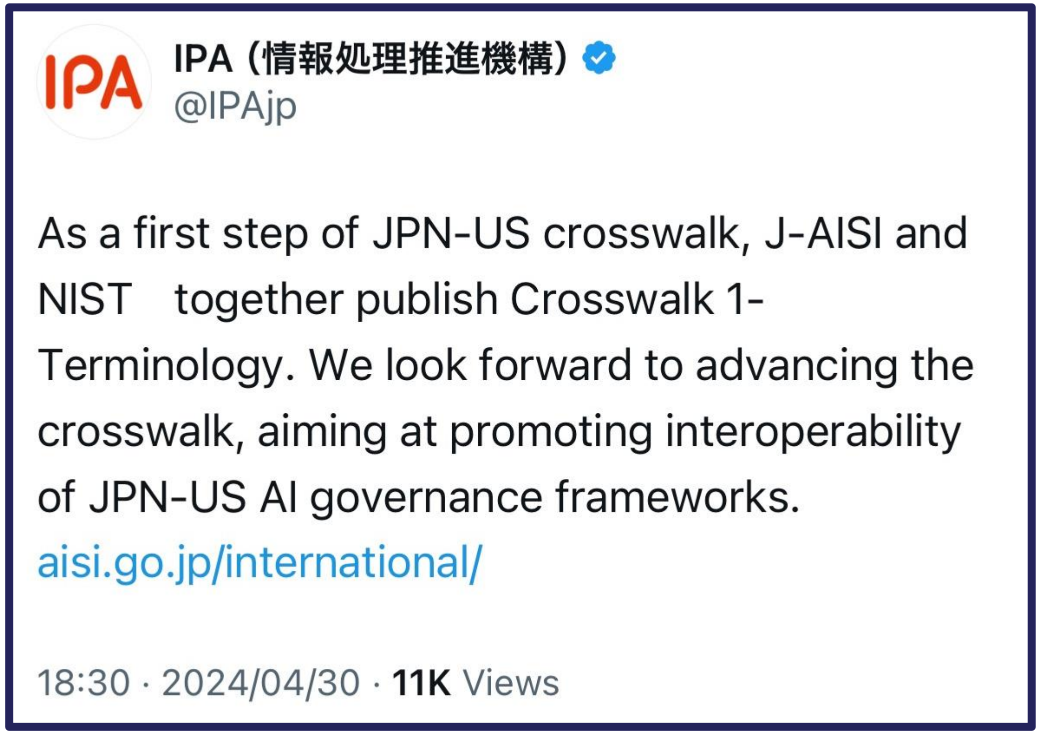
RMF	All	Developer	Provider	User
Govern				
Map				
Measure				
Manage				

U.S. point of view

→ Essential item

Overview of Japan-U.S. Crosswalk

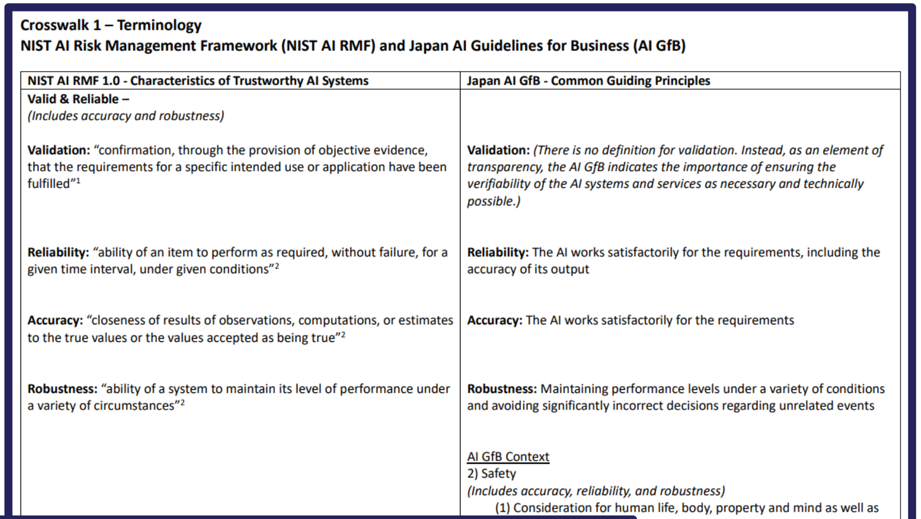
- Crosswalk 1 (Released on April 30th)
  Crosswalk1.pdf
- Crosswalk 2 (Released on September 18th)
  Crosswalk2.pdf



IPA (情報処理推進機構)
@IPAjp

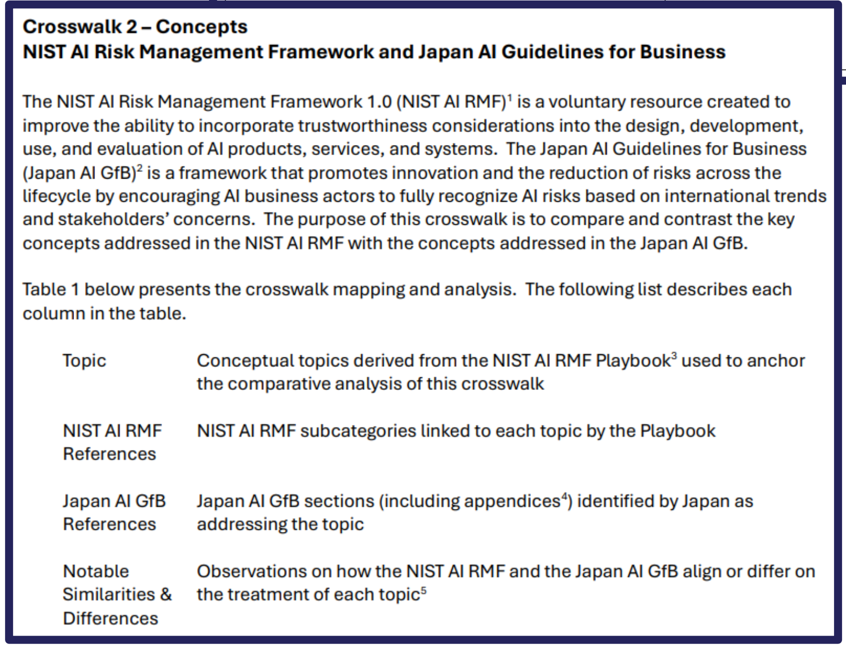
As a first step of JPN-US crosswalk, J-AISI and NIST together publish Crosswalk 1-Terminology. We look forward to advancing the crosswalk, aiming at promoting interoperability of JPN-US AI governance frameworks. aisi.go.jp/international/

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]	**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

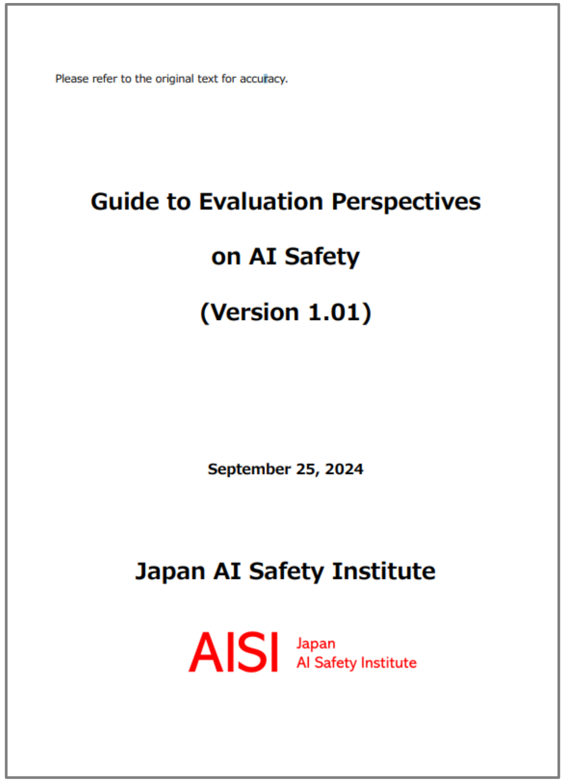
The NIST AI Risk Management Framework 1.0 (NIST AI RMF)[1] is a voluntary resource created to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. The Japan AI Guidelines for Business (Japan AI GfB)[2] 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

- 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.



Please refer to the original text for accuracy.

Guide to Evaluation Perspectives

on AI Safety

(Version 1.01)

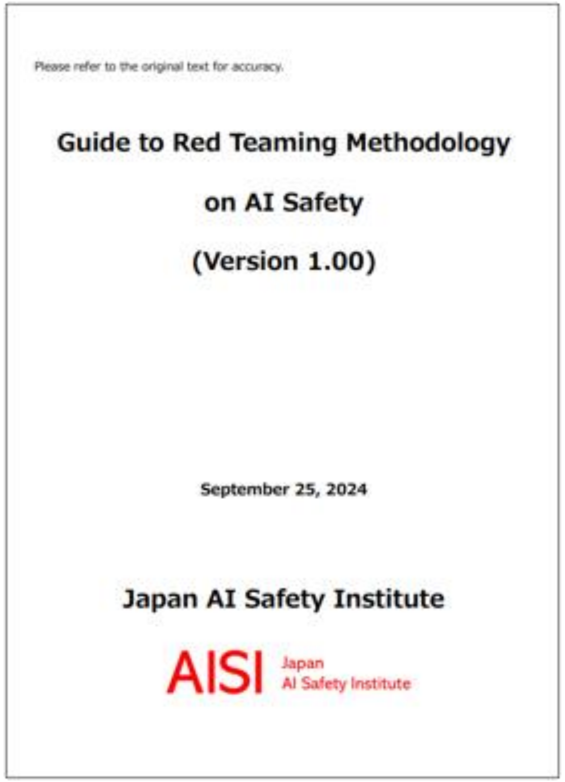
September 25, 2024

Japan AI Safety Institute

AISI Japan AI Safety Institute

Red Teaming Methodology Guide

- 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.



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

Guide to Red Teaming Methodology on AI Safety.pdf

International Cooperation

- 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 Fransico, 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
22 MAY 2024
Hosted by the Republic of Korea
Governance to Ensure Innovation,
Inclusivity and Safety of AI

AI Global Forum,
held in conjunction with AI Seoul Summit

United Nations
Summit of the Future

Plans for Future Initiatives

- 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
  - 2025 February 10-11(Paris)   Artificial Intelligence Action Summit

AISI

Japan AI Safety Institute