AI Safety Institute (AISI)

AISI Japan
AI Safety Institute

Background

- 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

- **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 activities to be harmonized with related domestic and international organizations.**
    - Response to rapidly and globally advancing technologies.

Role and Scope of AISI

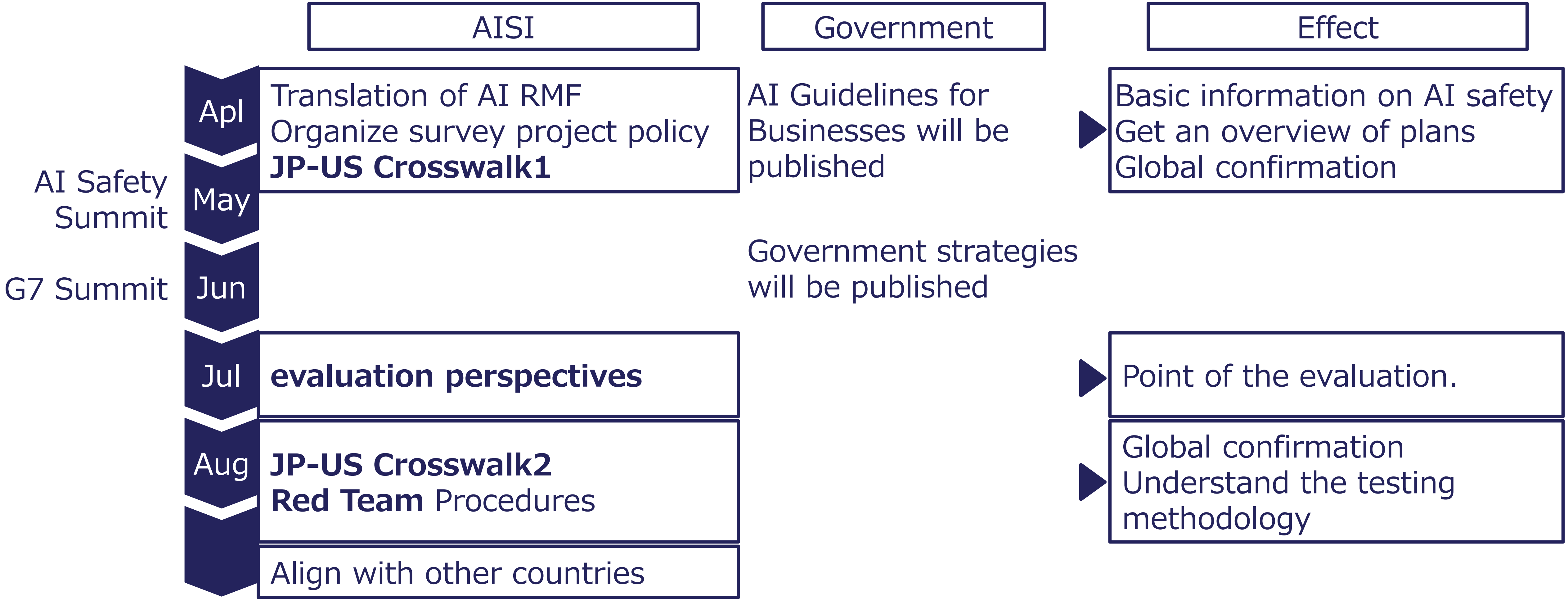
- Role
  - **AISI supports the government** by conducting surveys on AI safety, examining evaluation methods, and creating standards.
  - **As a hub for AI safety in Japan**, AISI will consolidate the latest information in industry and academia, and promote collaboration among related companies and organizations.
  - **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

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.)

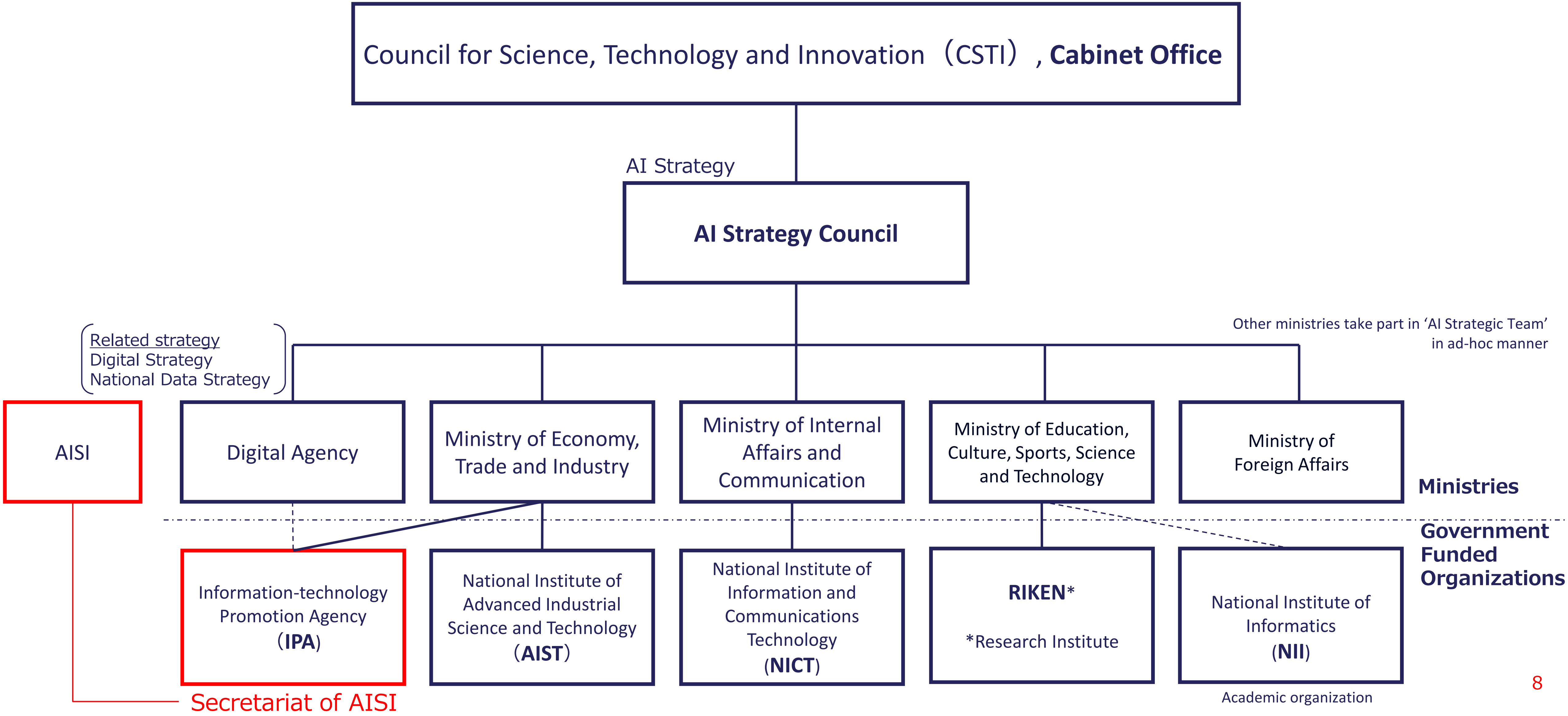
Immediate Activities and Deliverables



		AISI	Government	Effect
	Apl	Translation of AI RMF Organize survey project policy **JP-US Crosswalk1**	AI Guidelines for Businesses will be published	Basic information on AI safety Get an overview of plans Global confirmation
AI Safety Summit	May			
G7 Summit	Jun		Government strategies will be published	
	Jul	**evaluation perspectives**		Point of the evaluation.
	Aug	**JP-US Crosswalk2** **Red Team** Procedures		Global confirmation Understand the testing methodology
		Align with other countries		

Our team

AI Related Government organization



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

AI Strategy

AI Strategy Council

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

Related strategy
Digital Strategy
National Data Strategy

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

Government Funded Organizations

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

Secretariat of AISI

Executive Team

Executive Director

Akiko Murakami

Deputy Executive Director
Secretary General

Kenji Hiramoto

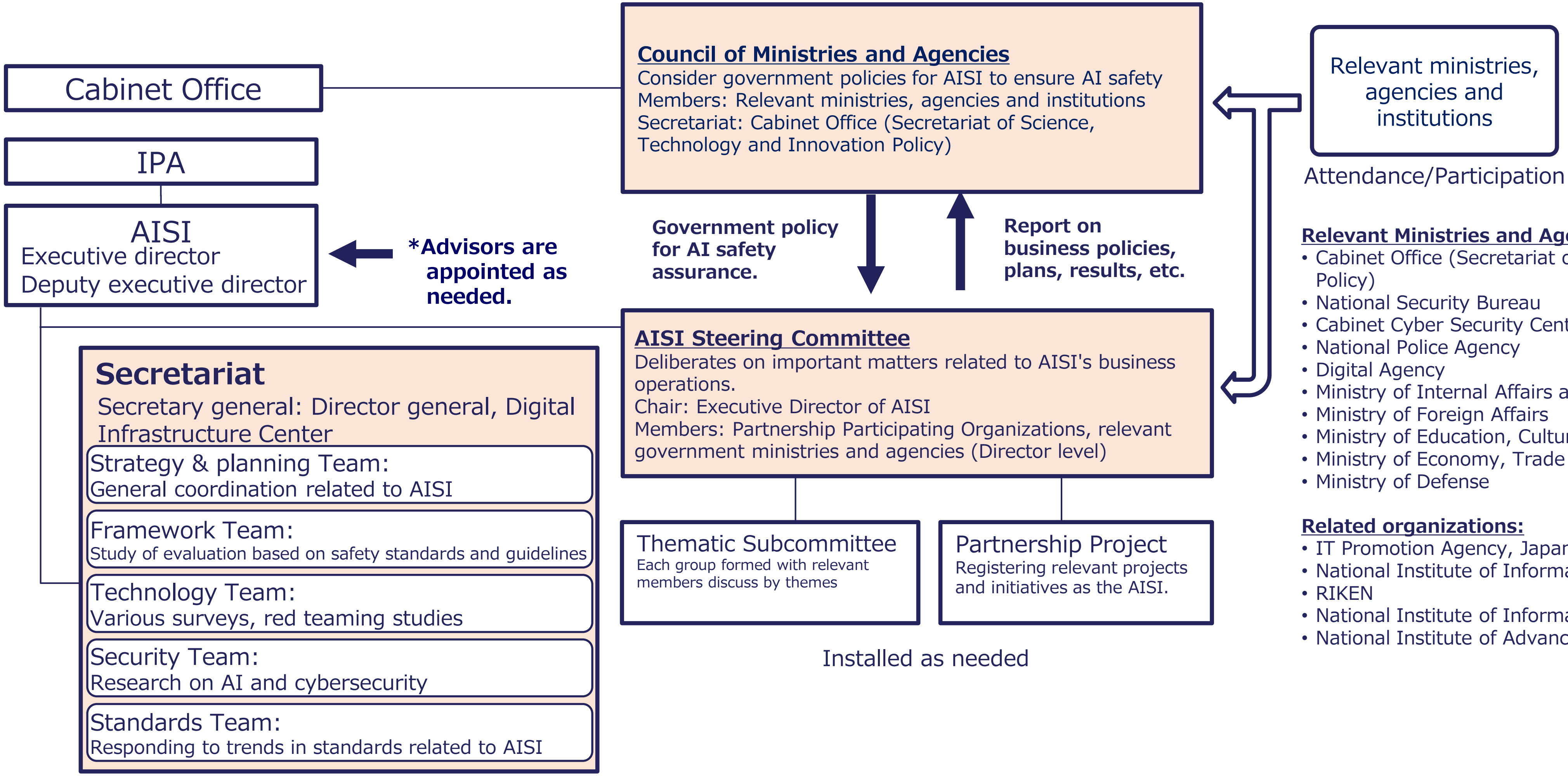
Deputy Executive Director

Hideyuki Teraoka

AISI Structures

AISI Japan AI Safety Institute

- "Council of Ministries and Agencies", set up in Cabinet Office, deliberates on the important matters of AISI.
- The "AISI Steering Committee" within AISI reports to the Council (to be held once a month). Under the Steering Committee, "thematic subcommittees" and "partnership projects" will be installed as necessary.
- As the secretariat of AISI, five teams were formed within the IPA Digital Infrastructure Center.



Cabinet Office

IPA

AISI
Executive director
Deputy executive director

***Advisors are appointed as needed.**

Council of Ministries and Agencies
Consider government policies for AISI to ensure AI safety
Members: Relevant ministries, agencies and institutions
Secretariat: Cabinet Office (Secretariat of Science, Technology and Innovation Policy)

Relevant ministries, agencies and institutions

Attendance/Participation

Government policy for AI safety assurance.

Report on business policies, plans, results, etc.

AISI Steering Committee
Deliberates on important matters related to AISI's business operations.
Chair: Executive Director of AISI
Members: Partnership Participating Organizations, relevant government ministries and agencies (Director level)

Thematic Subcommittee
Each group formed with relevant members discuss by themes

Partnership Project
Registering relevant projects and initiatives as the AISI.

Installed as needed

Secretariat
Secretary general: Director general, Digital Infrastructure Center

Strategy & planning Team:
General coordination related to AISI

Framework Team:
Study of evaluation based on safety standards and guidelines

Technology Team:
Various surveys, red teaming studies

Security Team:
Research on AI and cybersecurity

Standards Team:
Responding to trends in standards related to AISI

Relevant Ministries and Agencies:

- Cabinet Office (Secretariat of Science, Technology and Innovation Policy)
- National Security Bureau
- Cabinet Cyber Security Center
- 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

- **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**

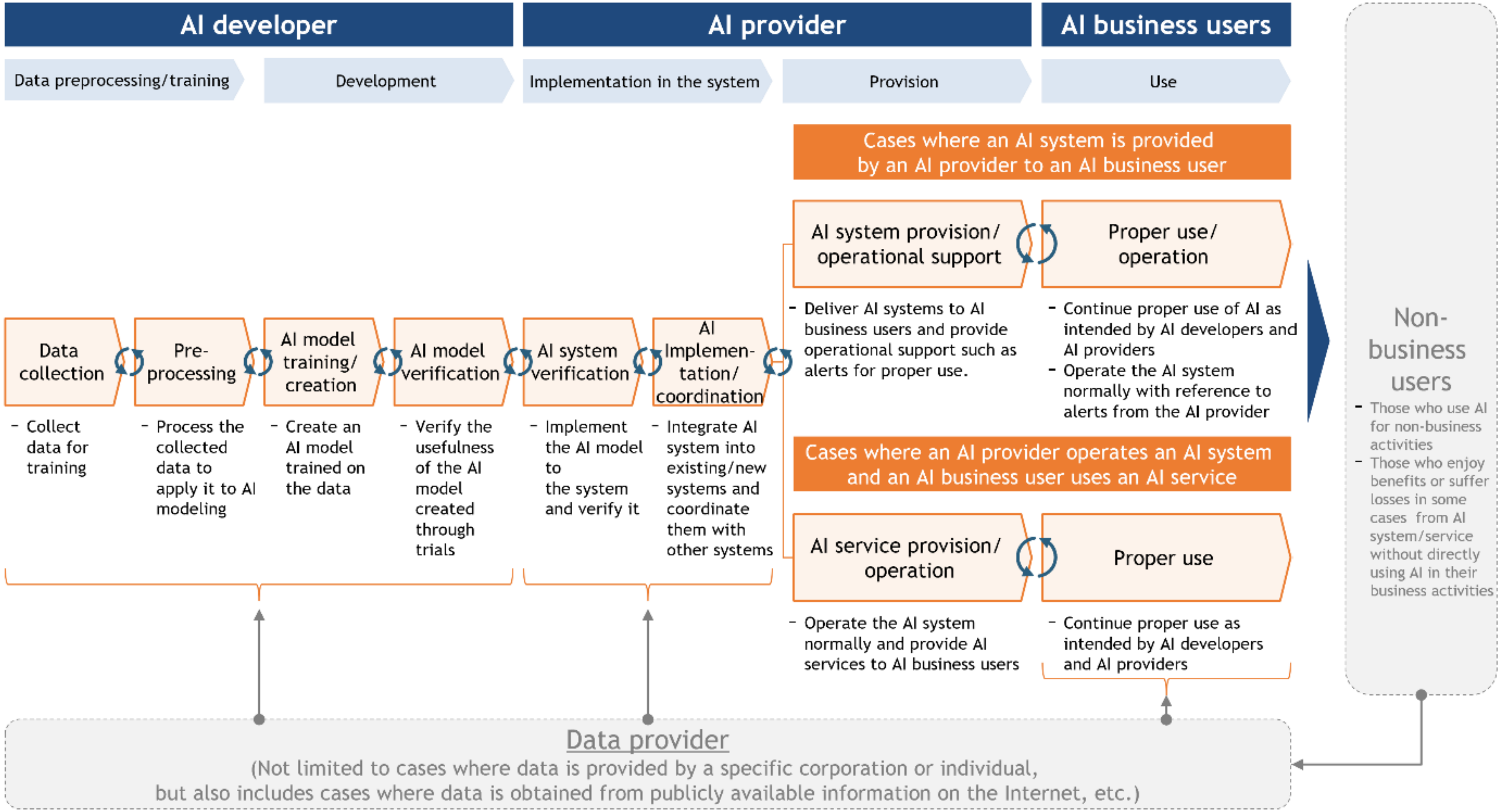
Items to be implemented at AISI

- **Technology**
  - Organizing **Red Team** Implementation Methodology
  - Collet 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 Implementation/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

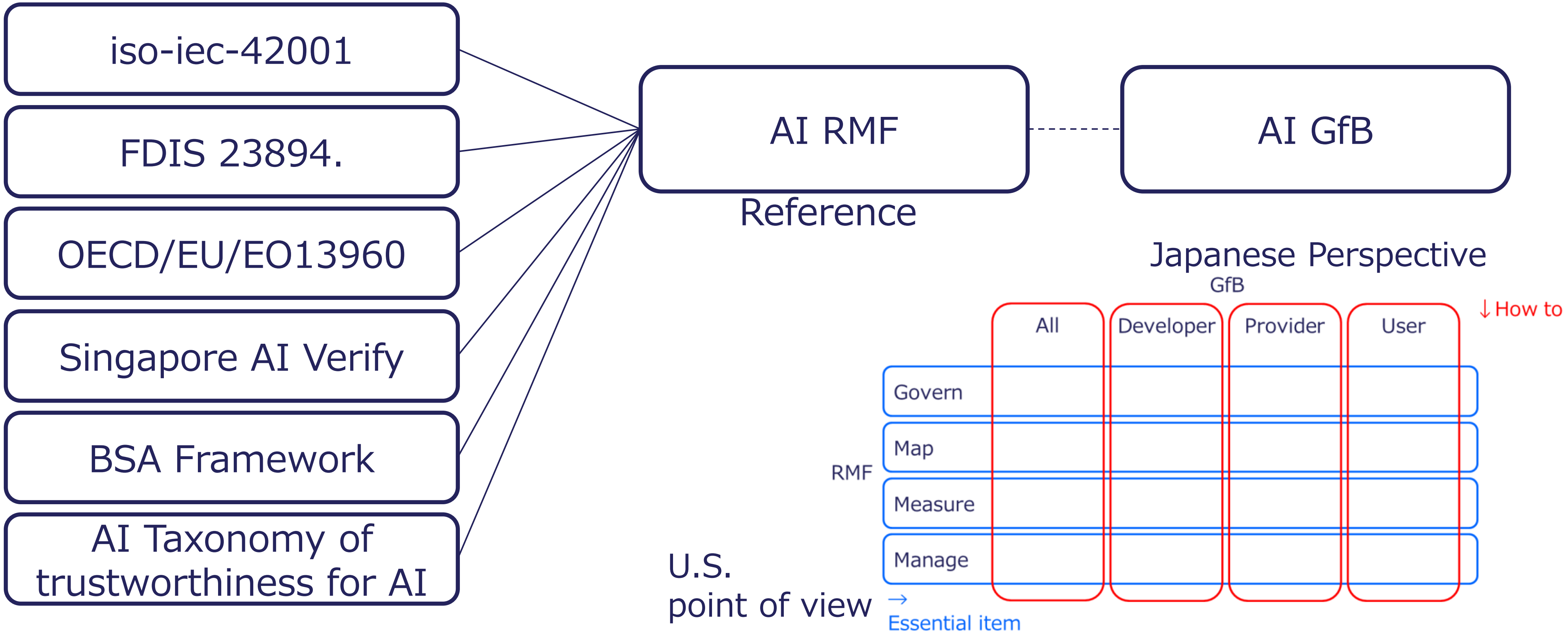
Non-business users
- Those who use AI for non-business activities
- Those who enjoy benefits or suffer losses in some cases from AI system/service without directly using AI in their business activities

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

Japanese Perspective

GfB

↓How to

RMF		All	Developer	Provider	User
	Govern				
	Map				
	Measure				
	Manage				

U.S.
point of view

→
Essential item

AISI

Japan AI Safety Institute