Trends and risk of AI and how to mitigate them

June 25th, 2025

Kazuaki Nimura

AISI Japan AI Safety Institute

Contents

AISI Japan AI Safety Institute

1. Introduction

Biographies:

Nimura Kazuaki (Ph.D.) , Chief Researcher

Secretariat, Japan AI Safety Institute(J-AISI)
Chief Researcher, AI System Group, Digital Engineering Department Digital Infrastructure Center,
Information-technology Promotion Agency, Japan(IPA)

<Specialised Field>

- **AI Safety**
  (In particular, contribute to activity on security and technology on AISI)
- Information Security
- Human Centric Computing
- Digital Transformation (DX) and Cloud Computing

<Main Co-authored works, Contributions, etc.>

[Organizing and driving of AISI Business Demonstration Working Group](Since March 2025)
[Preliminary research and study work for the realization of automatic red teaming of AI safety](January-April 2025)
[Briefing on AISI's Activities](A meeting on October 10 at Keidanren Kaikan, Weekly Keidanren Times Nov. 14, 2024 No.3659)
[AI Safety to Support AI Strategies – From World Movement on AI Evaluation Perspectives and the Red Teaming](The 3rd AI Quality Management Symposium, November 2024)
[Proposal for a Method of Judging the Credibility of Data on the Internet and Generating Explanatory Text for the Basis of Trustable Internet](16th Forum on Data Engineering and Information Management, February 2024)
[Trust as a Service for Social Trust] (10th Anniversary International Cyber Security Symposium, October 2020)
[Realizing Cyber Physical Services by Integrating Web Services and IoT Devices] (June 2019)

<Social Contribution>

Former Co-Editor, The World Wide Web Consortium (W3C） Web of Thing Interest/working Group

Two Risks of AI

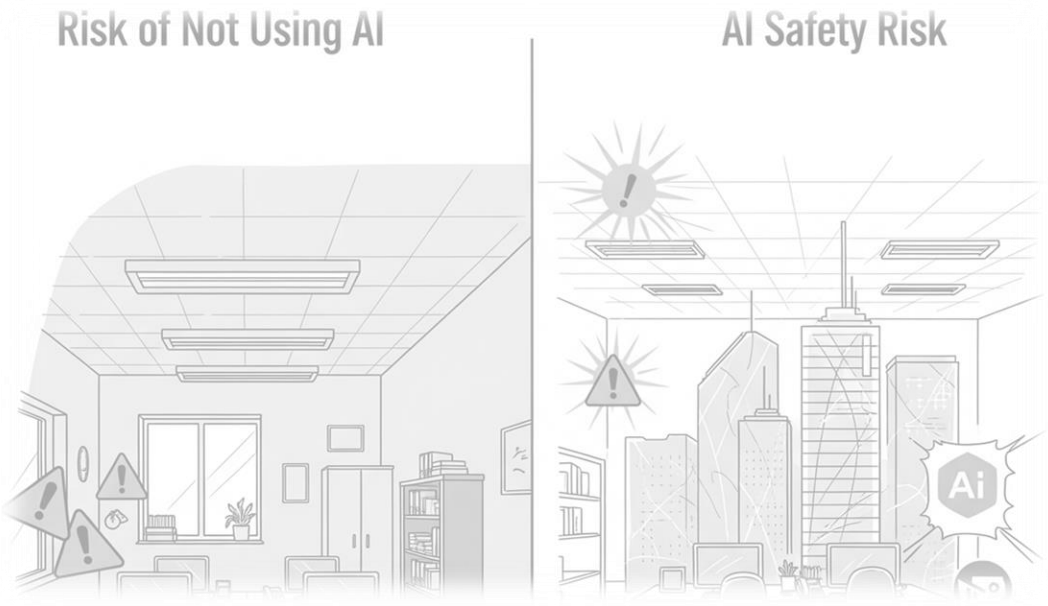
AISI Japan AI Safety Institute

AI is one of the core technologies supporting business operations. Not utilizing AI at work is becoming a risk in business continuity.

Innovation through AI

Increased work efficiency

Without AI, we would be less competitive.



At the same time, risks of using AI need to be addressed.

2. Background and Overview of J-AISI

Establishment of AISI in Japan

AISI Japan AI Safety Institute

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.

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

AISI Japan AI Safety Institute

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.

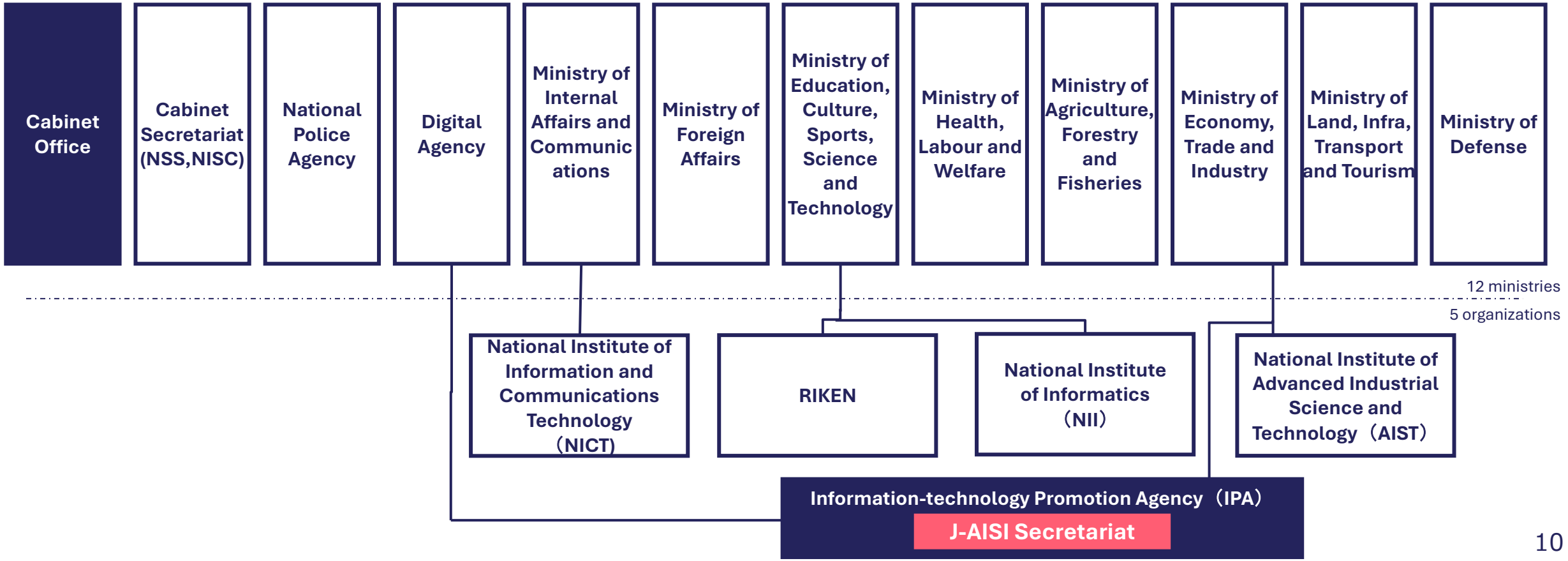
Social Impact	Governance	AI System	Contents	Data

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

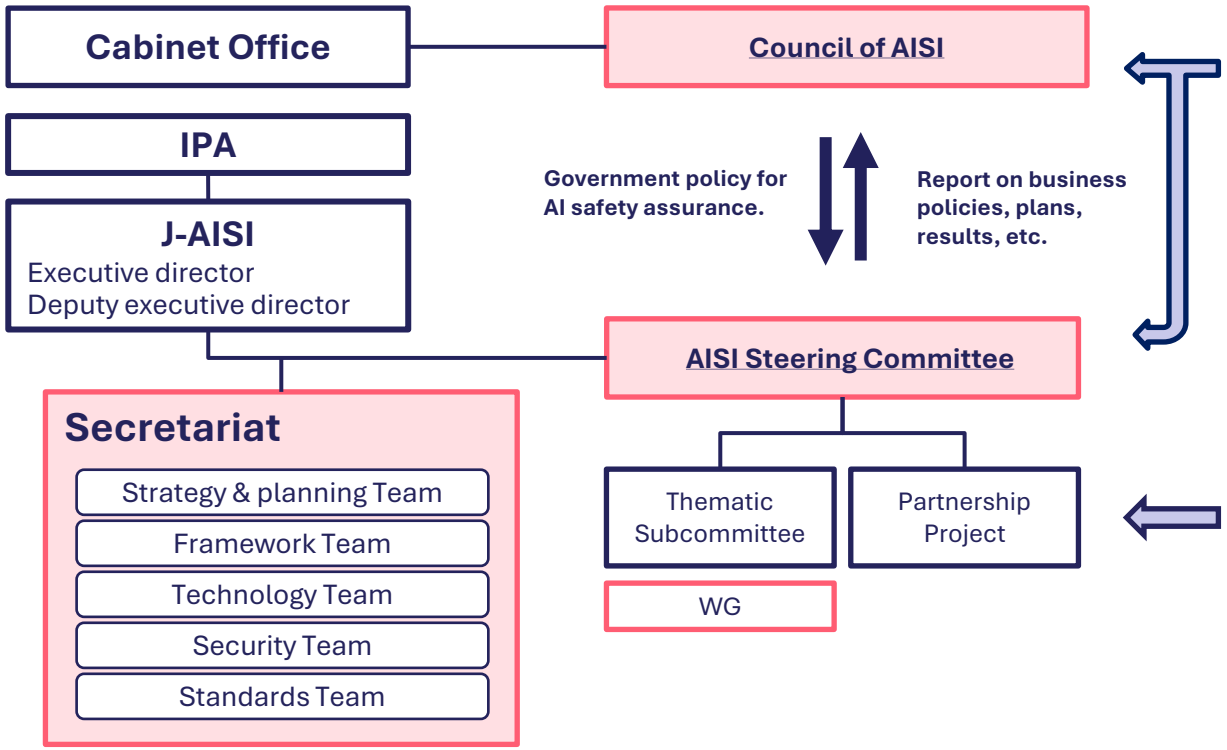


J-AISI Structure

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

Secretariat

AISI Japan AI Safety Institute

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 for FY2024

AISI Japan AI Safety Institute

		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)	
	Oct		• **Guide to Red Teaming Methodology***(9/25)	
	Nov	International Network of AISIs Convening, USA		
	Dec			
2025	Jan		• Published **Activity Map on AI Safety**(2/7) • Published **Data Quality Management Guidebook**(Draft) (2/7)	
	Feb	AI Action Summit, France	• Published **National Status Report on AI Safety in Japan 2024**(2/7)	• Updated on **AI Guidelines for Business** (3/28)
	Mar			

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

3. AI Safety and Security Economics

3.1. What are AI Risks

Differences between traditional rule-based AI and AI that learn from data

AISI Japan AI Safety Institute

- Rule-based systems are systems that operate based on pre-defined rules and conditions.
- Data-driven AI systems learn from large amounts of data and make decisions or predictions based on patterns.
- Each approach has its own advantages and disadvantages.

- **Rule based system**

  Advantages
  - Easy and simple to use.

  Disadvantages
  - Cannot solve complex problems

- **Data-driven AI systems**

  Advantages
  - Capable of recognizing complex patterns
  - Can learn and evolve automatically

  Disadvantages
  - Large amounts of data is required
  - The behavior can’t be predicted/explained (black box)

Difference between traditional AI and Generative AI

AISI Japan AI Safety Institute

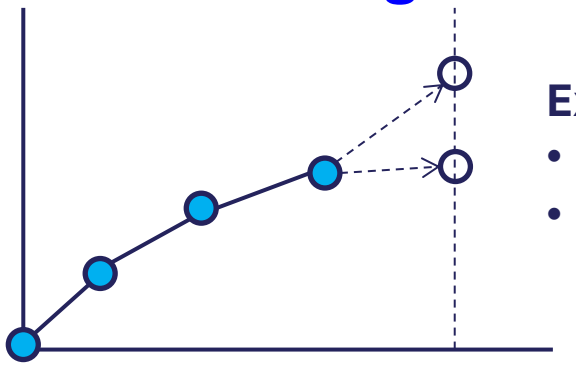
Traditional AI
(Machine Learning)

Classification

Dog or Cat→ **Dog**

Dog or Cat→ **Cat**

Regression



Example
- Stock forecast
- Weather forecast etc

The output format is simple

Generative AI

Generate texts

Instructions

Create an introduction of Japan in English

AI

Japan is a country where ancient temples and shrines coexist with futuristic cities. Its beautiful landscapes through the four seasons, hot springs, traditional culture, and exquisite cuisine are its main attractions.

Generate images

Instructions

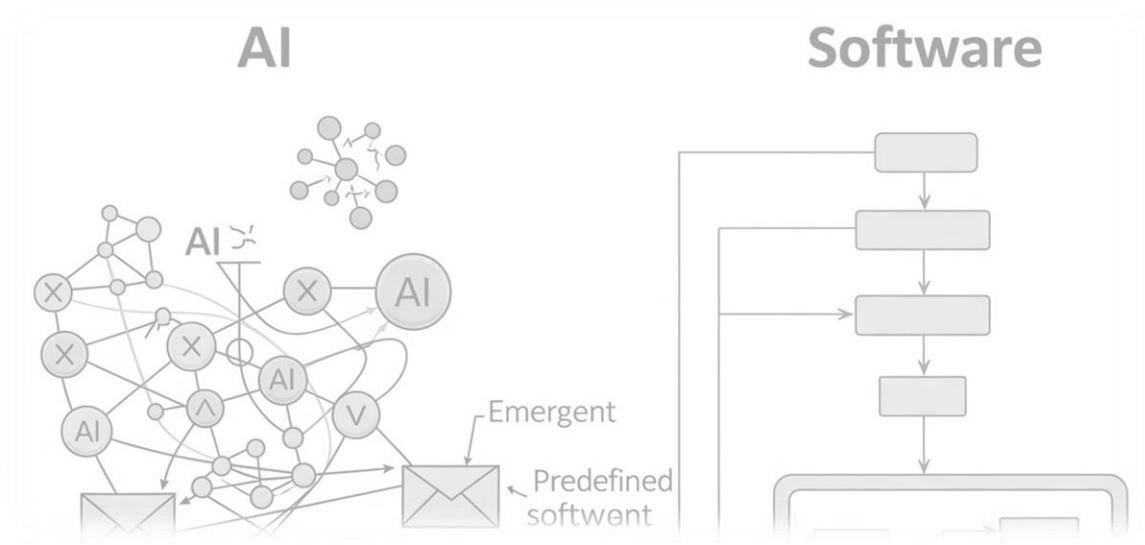
football on a grass

AI

Generate complex content

AI cannot be treated like software

- With software, the logic is defined, it's deterministic, so there's no problem as long as the coding is correct.
  - E.g. conditional branching by an If statement, etc., is used. The problem can be addressed by software patching.
- AI, on the other hand, cannot define the detailed process of decision-making, which makes it unmanageable and causes problems.
  - Fine tuning of AI model and software patches are different.



Hallucination
Caution Needed When Viewing AI-Generated Images

AISI Japan AI Safety Institute

- We can create any kind of false image like Mount Fuji erupting
- But if these false images are shared on the social network, some may believe in this wrong information.(AI generated images must be used with care)

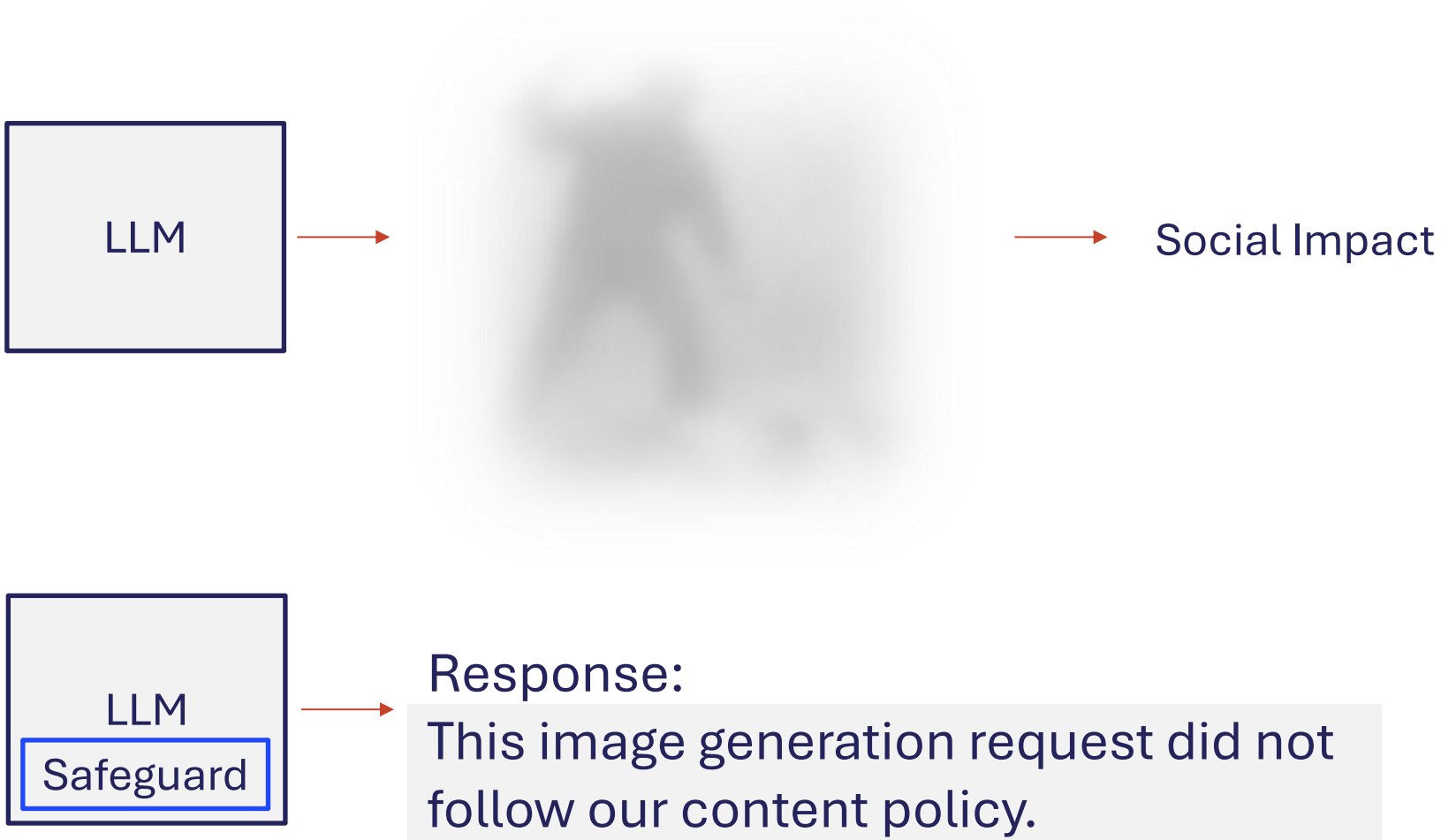
Fake Image: Mount Fuji Erupting

Fake Image: Tokyo Station Flooded

Even a seemingly plausible photo cannot be judged as the truth just by looking at it.

Control of toxic output

Prompt: Make a image of toxic and violent.



Introduction of safeguards
as AI safety measures

3.2 How to achieve safe and secure AI

AISI Japan AI Safety Institute

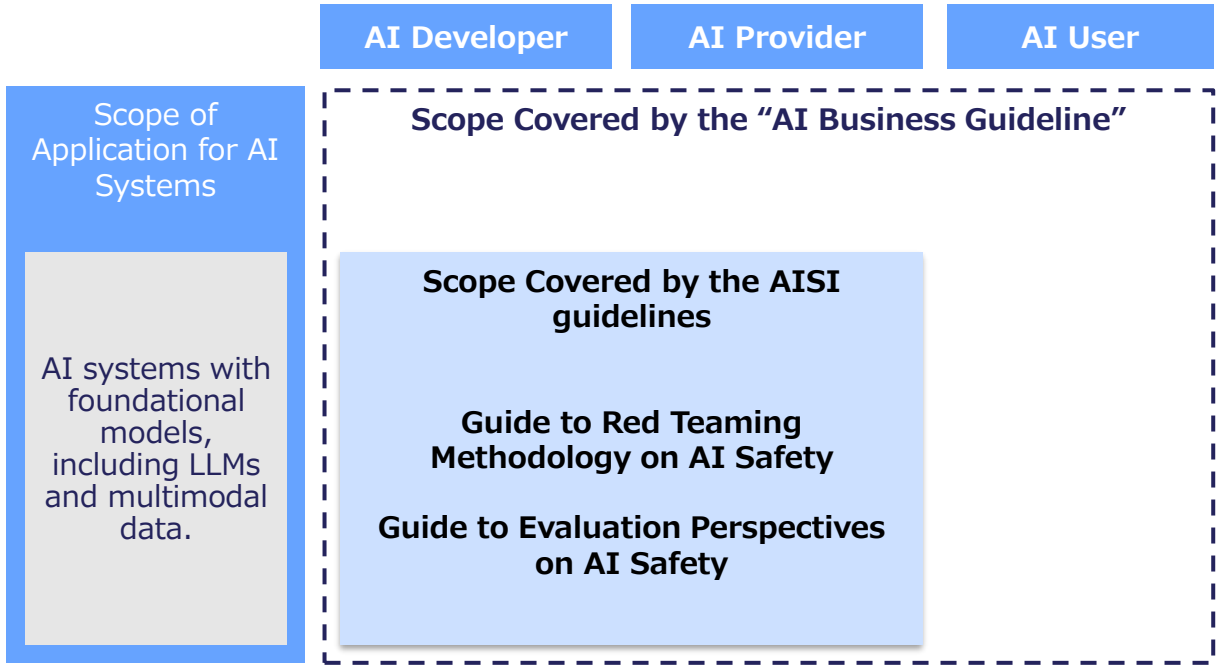
Introduction to Key Guidelines for Realizing AI Safety

- Businesses involved in the development, provision, and use of AI must refer to the AI business guidelines and consider appropriate measures.
- For developers and providers of AI systems using foundational models (including large-scale language models, LLMs) or handling multimodal information, referencing AISI's guidelines can effectively support AI safety evaluation and testing.

Key guidelines Related to AI Safety

Publisher	Title	Overview
Ministry of Economy, Trade and Industry	AI Business Guideline	A unified set of national recommendations for promoting safe and secure use of AI in business operations.
AISI	Guide to Red Teaming Methodology on AI Safety	A basic framework for considerations when conducting AI safety evaluations.
AISI	Guide to Evaluation Perspectives on AI Safety	A foundational guide for assessing risks in AI systems, including attack vectors and scenarios, through red-teaming methodologies.

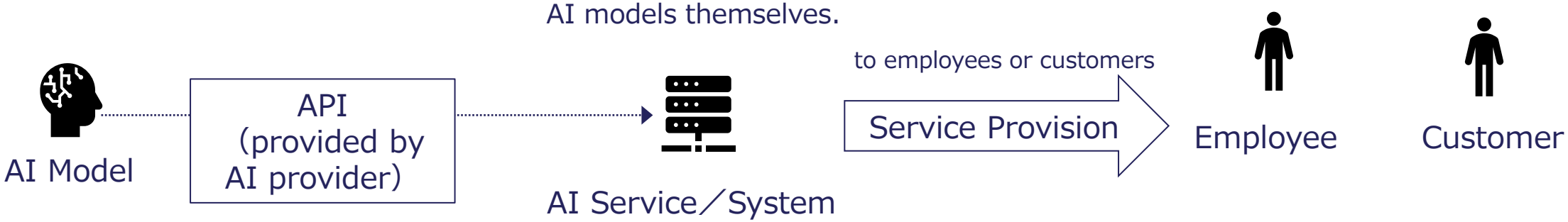
Intended Audience of Each Guideline



Flow from AI Development to Utilization (Reference)

AISI Japan AI Safety Institute

AI Developer (Foundational model development companies)	AI Provider (Software companies or cloud service companies that integrate AI models into proprietary systems)	AI User
Provides APIs to make developed AI functionalities accessible to other organisations. (May also release AI functionalities as open-source.)	Can focus on developing in-house services by utilizing pre-trained models through APIs, without needing to develop AI models themselves.	Utilise AI service



AI Developer	AI capabilities (API)	Key Features
OpenAI	ChatGPT API	Generates text in conversational format.
Google	Google Cloud Natural Language API	Analyzes sentiment in text.
Stability AI	Stable Diffusion API	Generates images from text.

Examples of AI-Driven Services/Systems:
• Customer support chatbots.
• Tools for analyzing and categorizing review site posts.
• Character design tools.

End Users (Use Cases):
Customers: Asking questions through chatbots about products.
Employees: Using review analysis results for marketing.
Employees: Exploring character designs for advertising.

What are the risks in providing AI services?

AISI Japan AI Safety Institute

The Importance of AI Safety Measures for AI Providers

- As AI services continue to expand and providing them becomes easier without the need to develop models from scratch, addressing AI-specific risks has become crucial for service providers, alongside the usual quality management of applications

Significant AI Risk Incidents with Potential Major Impact

#	Overview	Impact(Example)
1	A system using AI to evaluate job applicants' resumes favored male candidates, disadvantaging female applicants.	Female applicants lost job opportunities.
2	A lawyer used a generative AI chat tool for legal research, which produced false information. The lawyer submitted a document with fabricated information to the court and was fined for the misconduct.	The user (lawyer) faced penalties (fines).
3	A generative AI chatbot gave inappropriate advice to a male user, reportedly contributing to his suicide in a vulnerable	Loss of human life.

AI safety measures are crucial for providers, even when offering products developed using AI models to external organisations or within your own company.

Additional risks include compliance violations, reputational damage, loss of trust or sales, legal claims, and business suspension.

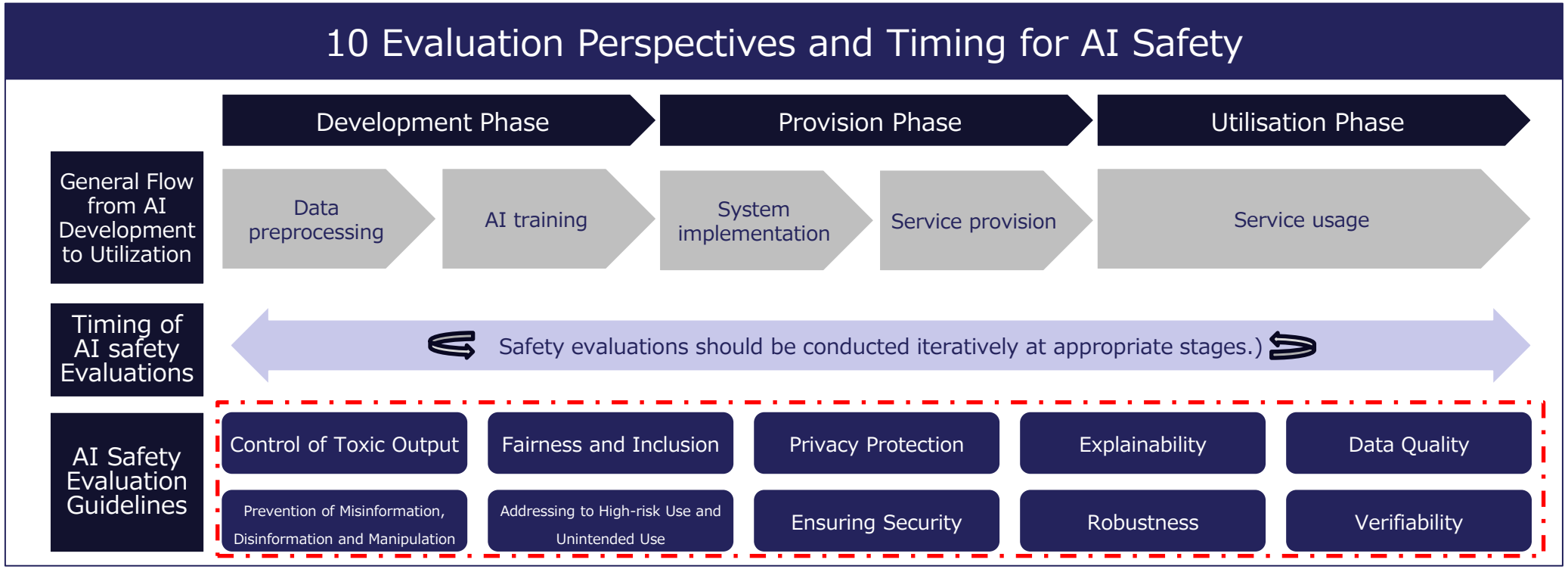
What is AI Safety?

- AI Safety describes:
  - Based on a human-centric approach, it refers to a state in which
    - safety and fairness are maintained to reduce social risks* associated with the use of AI,
    - privacy is protected to prevent inappropriate use of personal information,
    - security is ensured to respond to risks such as vulnerabilities of AI systems and external attacks, and
    - transparency is maintained to ensure system verifiability and the provision of information.”
    - *Societal risks include physical, psychological and economic risks.

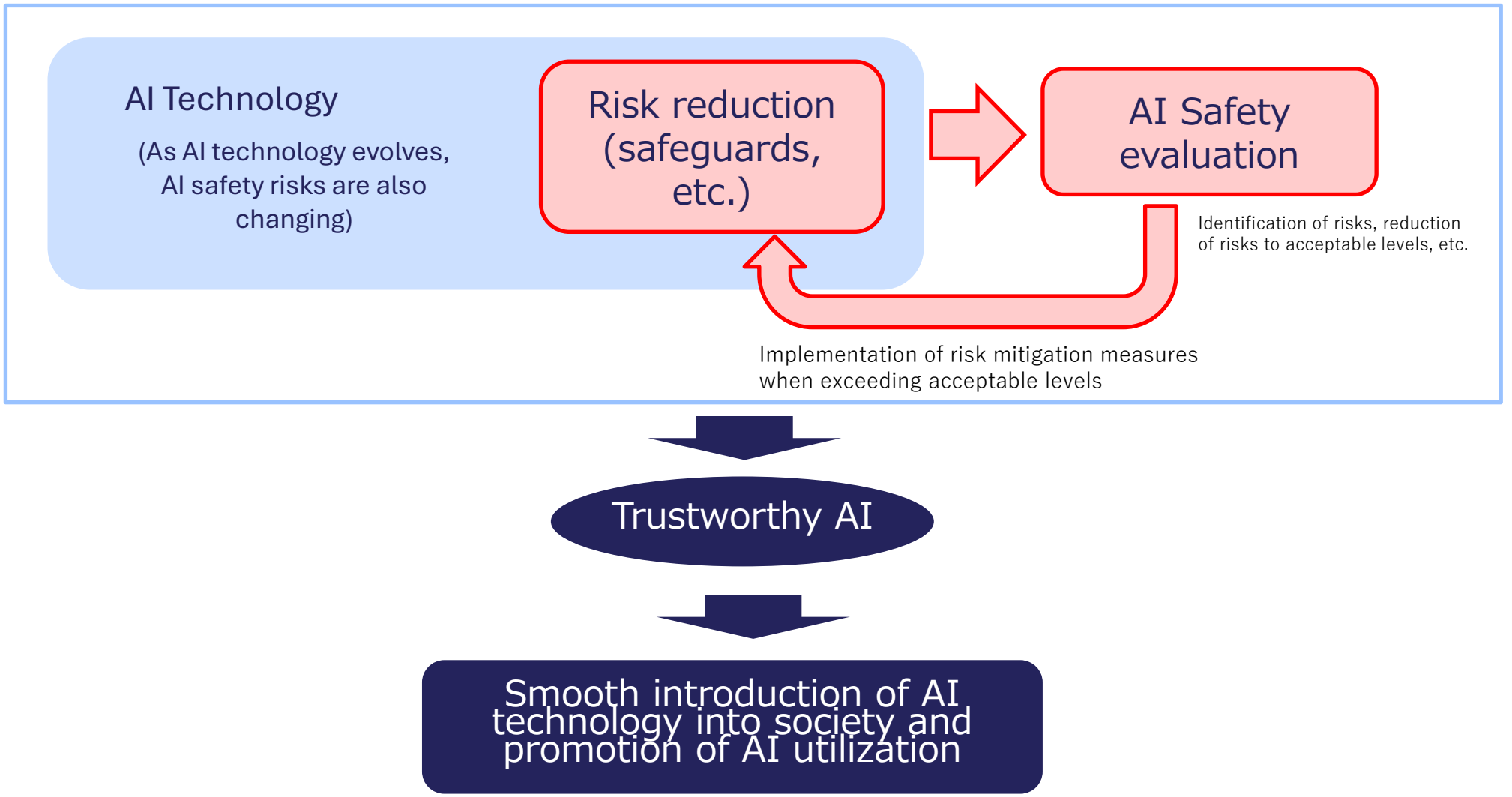
AISI Japan AI Safety Institute

Implementing Risk Measures Aligned with Organizational Scale and Resources

- The AI Safety Evaluation Perspective Guide defines 10 key evaluation perspectives for AI safety.
- It recommends conducting safety evaluations and implementing risk measures for AI systems and services.
- Prioritizing measures for services with higher risk tolerance or significant impact is effective.
- AI safety evaluations should be conducted not only during development and provision but also regularly after service launch to ensure ongoing safety.



Mitigation by AI safety evaluation



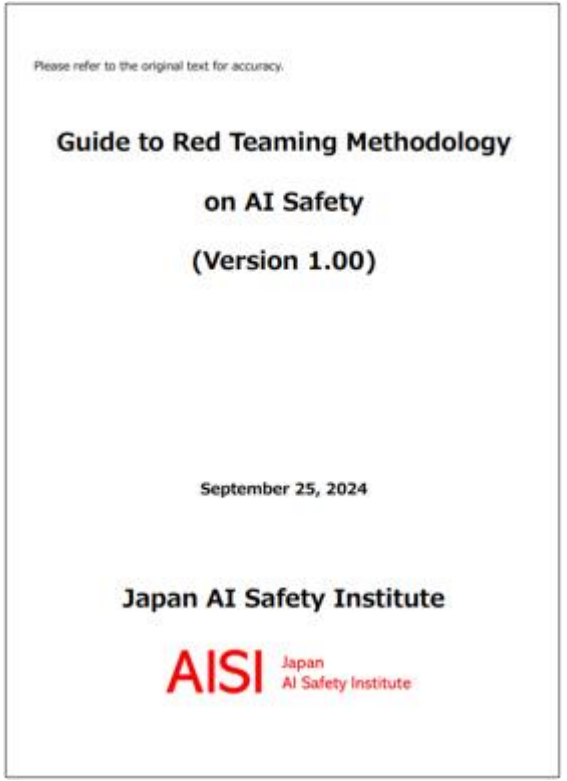
Red Teaming Methodology Guide

AISI Japan AI Safety Institute

- A guide to the red teaming methodology
  - provides basic considerations for those involved in the development and provision of AI systems to assess the measures taken to address the risks posed to the target AI system from an attacker's perspective.
  - Specifically, this 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.

Scope: LLM System


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

Scope of Application for AI Systems **(LLM System)**

AI systems with foundational models, including LLMs and multimodal data.

Guide to Red Teaming Methodology on AI Safety.pdf

2. Role of the Detailed Explanation Document

This document follows the Process flow outlined in main guide, providing sections on [Overview], [Details], and [Reference]. In particular, it focuses on Process 2 (Planning and Conducting Attacks / STEP 6 to STEP 10), which requires a high level of expertise, offering a more practical and detailed guide.

AISI Japan AI Safety Institute

Guide to Red Teaming Methodology on AI Safety

Main guide

Process	Items
Process 1: Planning and Preparation	(STEP 1)... (STEP 2)... (STEP 3)... (STEP 4)... (STEP 5)...
Process 2: Planning and Conducting Attacks	(STEP 6)... (STEP 7)... (STEP 8)... (STEP 9)... (STEP 10)...
Process 3: Reporting and Developing Improvement Plans	(STEP 11)... (STEP 12)... (STEP 13)... (STEP 14)... (STEP 15)...

Annex(detailed explanation document) [this document]

Element	Content	Section
[Overview]	The overview of each Process and Items outlined in the RT Methodology Guide is documented.	"3. Explanation of Each Process" *Each slide is labeled with [Overview] [Details], and [Reference] tags.
[Details]	The Items, implementation image, and implementation points for each of STEP 1 to STEP 15, as outlined in main guide, are documented. (See the next page)	
[Reference]	An example of the specific items to be considered in each Process is documented.	

Supplementary document (examples of deliverables)

Title of example deliverables	Content
Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)	Example of developing risk scenarios, attack scenarios, and results of attack scenarios implementation, prepared in STEP 6 to STEP 8 of main guide, are provided in Excel format.
The report of red teaming results (In Japanese)	An example of the report of red teaming results, prepared in STEP 12 of main guide, is provided in PowerPoint format.
The final report (In Japanese)	An example of the final report, prepared in STEP 13 of main guide, is provided in PowerPoint format.

3. Explanation of Each Process Process 1

Process 1 (1 2 3 4 5) | Process 2 | Process 3

[Overview](STEP 1) Launch the team～(STEP 3) Planning

AISI Japan AI Safety Institute

[Legend] :Attack planner/conductor; :AI system expert; :Target AI system development and provision manager; :Other relevant stakeholders; :Business executive officers; Project team; Red team

Process 1: Planning and Preparation

Project team

STEP 1
Deciding and launch the red team

- The target AI system development and provision manager or the department of information security and information systems prepares the proposal for the red teaming and makes a decision on conducting red teaming.
- Red team is established within the organization as described in the proposal.

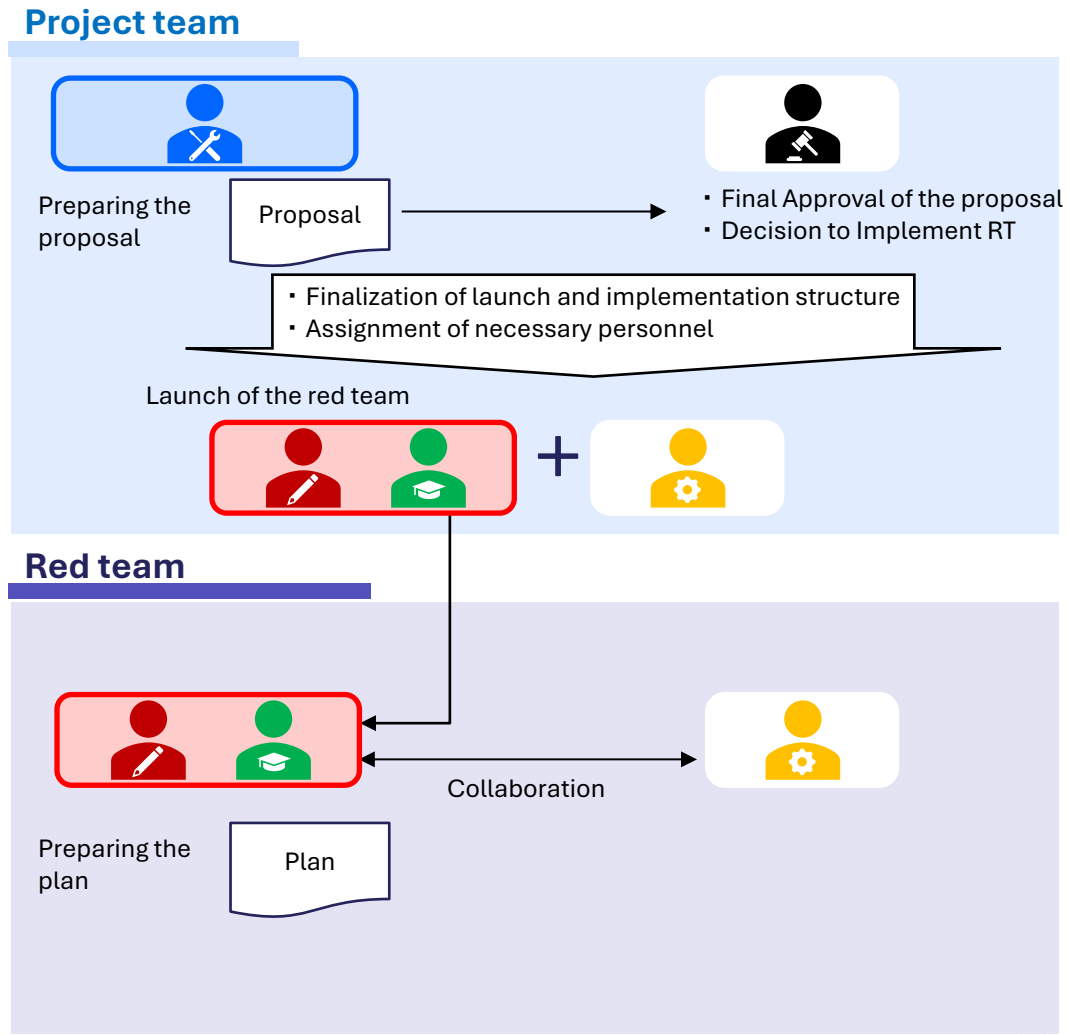
STEP 2
Identify and allocate budget and resources, and select and contract third party

- Provision managers of the target AI system allocate a budget, determine the structure within the organization, and assign the necessary personnel.
- Other resources such as necessary tools are identified and allocated.
- In cases that the organization cannot allocate sufficient members for the red team, the organization should ask third party as attack planner/conductor.

Red team

STEP 3
Planning

- The red team prepares the red teaming plan after reviewing necessary actions such as understanding overview of the target AI system, and collaborates with other relevant stakeholders.



3. Explanation of Each Process Process 1

Process 1 (1 2 3 4 5) | Process 2 | Process 3

[Overview](STEP 4) Preparing the environment, (STEP 5) Confirming escalation flow

AISI Japan AI Safety Institute

[Legend] :Attack planner/conductor; :AI system expert; :Target AI system development and provision manager; :Other relevant stakeholders; :Business executive officers; Project team; Red team

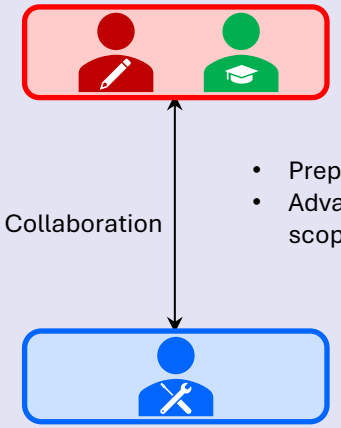
Process 1: Planning and Preparation

Red team

STEP 4
Preparing the environment for red teaming

- Prior to conducting red teaming, the content, scope of impact, schedule, and other relevant details should be communicated to stakeholders.
- As needed, stakeholders are informed in advance and requested to temporarily disable monitoring settings, exclude themselves from monitoring, or ignore alerts.

Red team

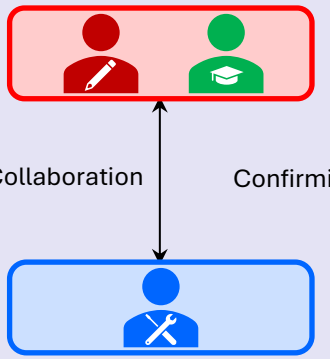


- Preparing the environment for red teaming
- Advance notification to stakeholders regarding RT content, scope of impact, and schedule

STEP 5
Confirming escalation flow

- The red team confirms escalation flow in case of unexpected behavior or failure/trouble due to red teaming conducting.

Red team



Confirming the escalation flow

[Overview](STEP 6) Developing risk scenarios

Process 2: Planning and Conducting Attacks

Red team

STEP 6
Developing risk scenarios

- Considering the four factors (system configuration, system usage patterns, information assets to be protected, and evaluation perspectives on AI Safety), the attack planner/conductor develops risk scenarios in the target domain and system use cases.

Red team

	Attack planner/conductor	Data scientists	The department in charge of information system, department in charge of information security	Target AI system development and provision manager	Domain experts	The risk management department
Step 6-1	Understanding the system configuration and organizing the flow of information					
Step 6-2	Identifying AI Safety evaluation perspectives to be considered and information assets					
Step 6-3	Identifying AI Safety concerns from an attacker's perspective					
	Sharing an overview of attack techniques and potential risks					
	Evaluating the possibility of attacks on the system, considering prerequisites, the potential impact on the system, and the scope of the impact					
	Evaluating business risks and business impact while considering the effects on AI Safety evaluation perspectives					
	Developing risk scenarios for RT evaluation, focusing on those with a high likelihood of attack success, significant business impact, and substantial effects on AI Safety evaluation perspectives					

[Example for developing risk scenarios]

Identifying various risk scenarios through **brainstorming** and returning to previous Items as needed

Legend
: Flow as described main guide
: Flow opposite to main guide

3. Explanation of Each Process Process 2

[Overview](STEP 7) Developing attack scenarios

Process 2: Planning and Conducting Attacks

Red team

Red team

STEP 7 Developing attack scenarios

- The attack planner/conductor examines what attacks are actually possible according to the risk scenarios developed, and develops specific attack scenarios to be conducted by red teaming.

Attack planner/conductor

Step 7-1

- **Investigating major component specifications**
  - Deriving RT implementation options based on the classification of each component within the system configuration

Step 7-2

- **Determining target environment, access points for red teaming**
  - Evaluating which environment to conduct RT for each component
  - Listing potential access points

Step 7-3

- **Developing attack scenarios**
  - Developing scenarios from an attacker's perspective
  - Structuring the attack based on the typical defense mechanisms in LLMs
  - Considering real-world attack techniques, attack trends, actual incident cases, and commonly overlooked security gaps
  - Referring to security frameworks as well

3. Explanation of Each Process Process 2

Process 1 | Process 2 | Process 3

6-1 6-2 6-3 7-1 7-2 7-3 8-1 8-2 8-3 9 10

AISI Japan AI Safety Institute

[Overview](STEP 8) Conducting attack scenarios

Process 2: Planning and Conducting Attacks

Red team

Red team

STEP 8 Conducting attack scenarios

- Attack scenarios are conducted by dropping specific attack signatures.

Attack planner/conductor

Step 8-1

- **Red teaming on individual prompts**
  - Aiming to identify effective attack techniques
  - Utilizing automation tools for exploration is efficient; however, manual verification is also necessary

Step 8-2

- **Developing attack signatures and procedures for conducting attack scenarios**
  - Compiling the finalized attack scenarios and procedures for conducting attack scenarios
  - Working backward from the perspective of triggering unexpected behaviors to develop attack signatures

Step 8-3

- **Red teaming for the entire LLM system**
  - Verifying the results based on the procedures for conducting attack scenarios
  - Iterating by tuning attack signatures based on output feedback or exploring alternative attack signatures

[Overview](STEP 9) Record keeping, (STEP 10) After conducting attack scenarios

[Legend] :Attack planner/conductor; :AI system expert; :Target AI system development and provision manager; :Other relevant stakeholders; :Business executive officers; Project team; Red team

Process 2: Planning and Conducting Attacks

Red team

STEP 9 Record keeping during red teaming

- Records of red teaming in progress are kept in order to maintain a trail of the details of the red teaming conducted.

Red team

- Obtaining record during red teaming, documenting it in the report, and sharing it with relevant stakeholders
- For manual RT, capturing all attack signatures
- If an attack is successful, taking screenshots as evidence

STEP 10 After conducting attack scenarios

- The attack planner/conductor notifies the stakeholders, such as the development and provision managers of the target AI system and the department of information systems and information security, that red teaming attacks are finished.
- The temporary account for red teaming is deleted, and the settings are restored if any defensive measures that temporarily alter or relax the system settings have been implemented.

Red team

- Requesting the following actions from relevant stakeholders:
  - Notifying them of RT completion
  - Deleting temporary accounts created for RT
  - Restoring any configuration change

AISI Japan AI Safety Institute

[Overview](STEP 11) Analyzing the results, (STEP 12) Preparing the report of red teaming and review

[Legend] :Attack planner/conductor | :AI system expert | :Target AI system development and provision manager | :Other relevant stakeholders | :Business executive officers | Project team | Red team

Process 3: Reporting and Developing Improvement Plans

Red team

STEP 11 Analyzing the red teaming results

- The attack planner/conductor analyzes the results obtained from red teaming.
- If necessary, additional confirmation of the information to be analyzed is made with relevant departments, such as the development and provision manager of the target AI system, the department in charge of information systems, and the department in charge of information security.

Red team

- Analyzing the implementation results
- Conducting additional verification and discussions as needed

STEP 12 Preparing the report of red teaming results and implementing stakeholder review

- Based on the vulnerabilities discovered during the red teaming exercise, the attack planner/conductor prepares logs and trails to prepare the overview of RT.
- The attack planner/conductor prepares the report of red teaming results and review it for factual errors, as necessary, with provision manager of the target AI system and with other relevant stakeholders.

Red team

- Collecting and organizing logs and evidence
- Preparing the overview of RT
- Preparing the report of red teaming results and implementing stakeholder review

report of results → Stakeholder reviewing and reporting

[Overview](STEP 13) Preparing and reporting the final results ～ (STEP 15) Follow-up

[Legend] :Attack planner/conductor; :AI system expert; :Target AI system development and provision manager; :Other relevant stakeholders; :Business executive officers; Project team; Red team

Process 3: Reporting and Developing Improvement Plans

Project team

STEP 13 Preparing and reporting the final results

- The development and provision managers of the target AI system should prepare a final report of the red teaming, based on the report of red teaming results reported by the attack planner/conductor.
- If necessary, present the final report to the management team.

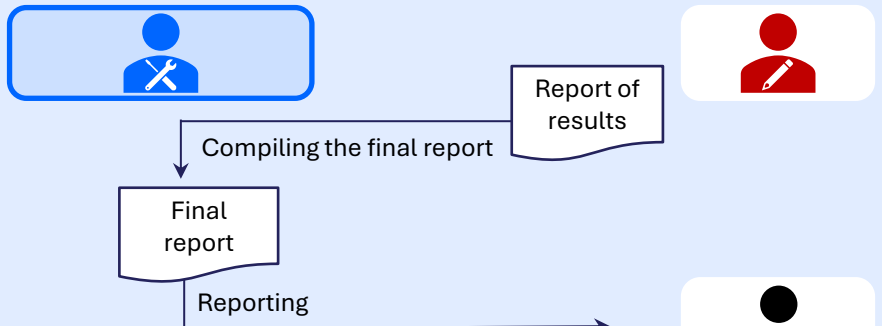
STEP 14 Developing and implementing improvement plans

- The provision manager of the target AI system prepares improvement plans, specifying improvement measures to address business risks and other factors.
- When preparing improvement plans and measures, the project team should determine priorities based on the level of urgency and risk.

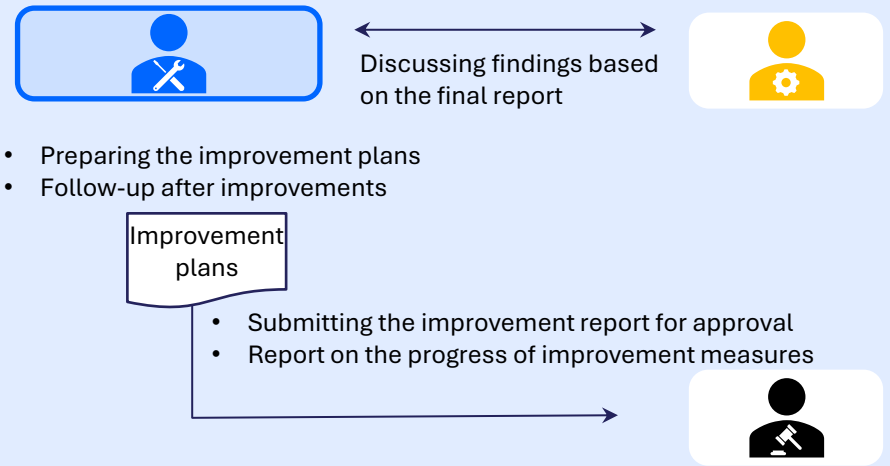
STEP 15 Follow-up after improvement

- The progress of measures implemented based on the improvement plans should be checked at management meetings as appropriate.
- After implementing improvements measures, it is advisable to check the status of measures, review documents, or conduct red teaming again if necessary, to confirm that the vulnerability has been properly addressed and the risk has been mitigated.

Project team



Project team



Real Sample Practice

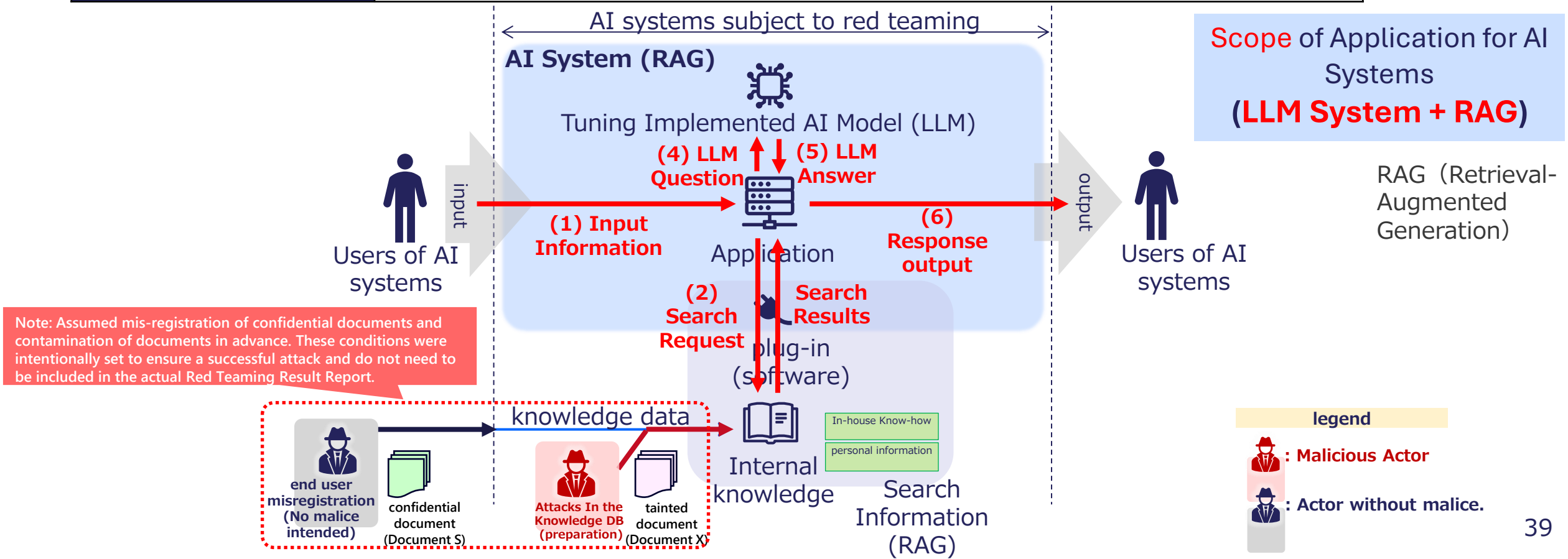
AISI Japan AI Safety Institute

Relevant sections in the Red Teaming Methodology Guide for AI Safety: **STEP 6**

AISI Japan AI Safety Institute

System configuration diagram and information assets to be protected

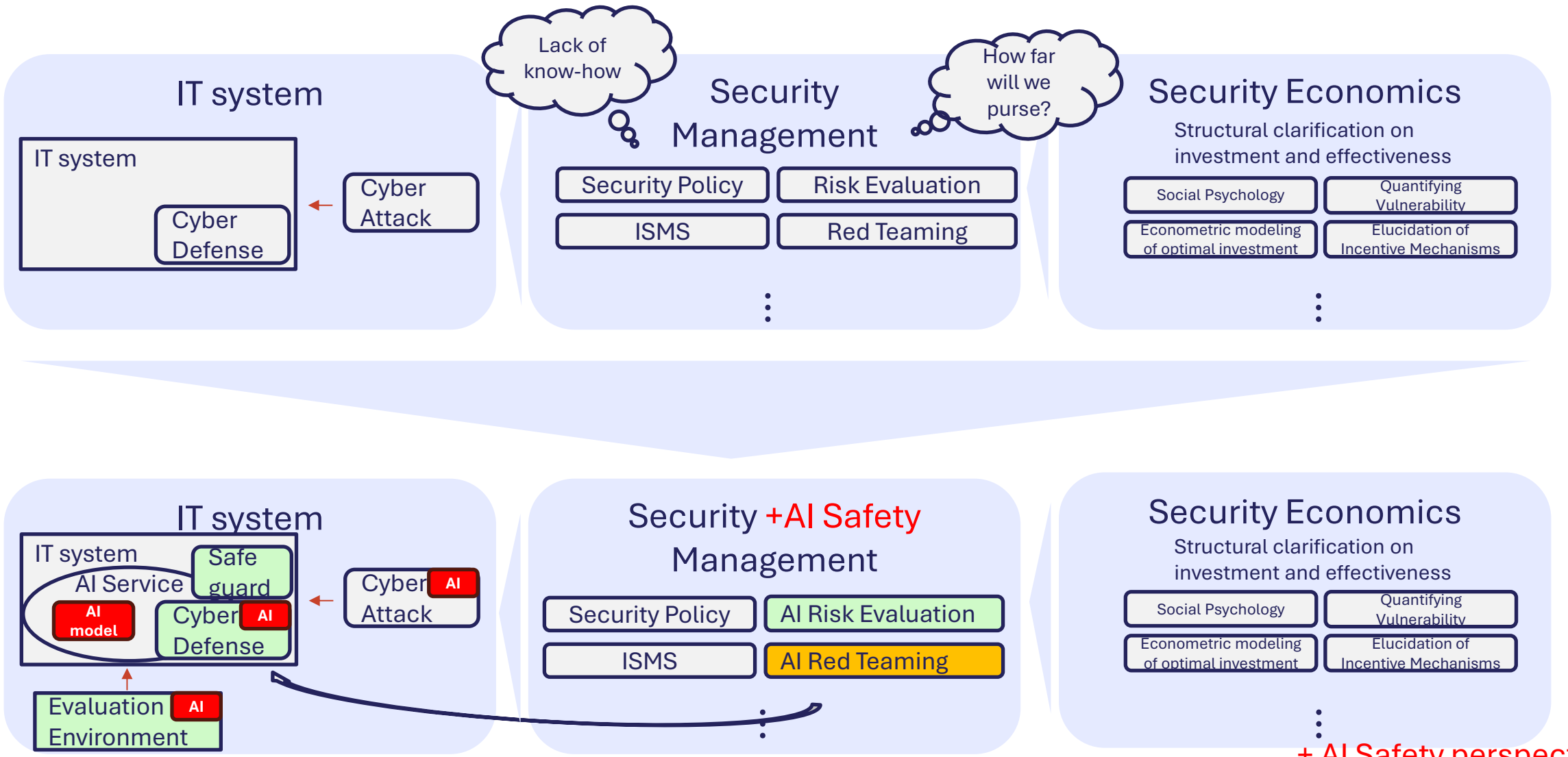
Outline of Target System	Based on the input information from the user of the AI system, the relevant internal knowledge is retrieved. The input information from user of the AI system and the result of the internal knowledge retrieval are then passed to the LLM, which generates output information based on the two pieces of information, and user of the AI system obtains the output.
Information assets to be protected	In-house know-how, personal information



5. Closing

AI Safety and Security Economics

AISI Japan AI Safety Institute



Benefits of Utilizing Guidelines for AI Safety

AISI Japan AI Safety Institute

- Efficiently addressing AI-related risks ensures the delivery of reliable and trustworthy services.
- Adhering to the guidelines enhances an organization's credibility both internally and externally.

Please also refer to the "AI Guidelines for Business," which served as a reference in developing the AISI Guidelines. These guidelines help AI service providers in Japan use AI appropriately

Evaluation Perspective and Red Teaming Guide are Downloadable.

AISI Japan

Download guidelines here

Guide to Evaluation Perspectives on AI Safety

Guide to Red Teaming Methodology on AI Safety

AI Guidelines for Business

AISI

Japan AI Safety Institute