Guide to Red Teaming Methodology on AI Safety (Version 1.10)

Summary

AI Safety Institute

(March 31, 2025)

AISI Japan AI Safety Institute

Table of Contents

1. Background and Purpose ・・・P.3

2. Document Writing Policy ・・・P.4

3. Structure of This Document ・・・P.5

4. Scope of Red Teaming ・・・P.6

5. Overview of Red Teaming ・・・P.7

6. Process of Red Teaming ・・・P.9
   - Planning and Preparation
   - Planning and Conducting Attacks
   - Reporting and Developing Improvement Plans

Revision of Guide to Red Teaming Methodology on AI Safety ・・・P.13

1. Background and Purpose

With the rapid proliferation of AI systems, AI Safety is becoming increasingly important. This document aims on presenting basic considerations for red teaming methodologies.

AISI Japan AI Safety Institute

Background

- While the development, provision, and use of AI systems are expected to promote innovation and solve social problems, concerns have arisen due to misuse and abuse of AI systems, inaccurate output, etc.
- There is a growing interest in so-called AI Safety both in Japan and abroad, and **as part of the AI Safety evaluation, the red teaming method, in particular, is being studied in many countries.**

Purpose

- The "Guide to Red Teaming Methodology on AI Safety" (hereafter referred to as "this document") provides **basic considerations for those involved in the development and provision of AI systems regarding red teaming methodologies to evaluate the risk countermeasures applied to the target AI system from an attacker's perspective**.
- This document presents items that are considered important for conducting red teaming at this stage, taking into consideration domestic and international studies, prior cases, and international alignment.

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.”

*Social risks include physical, psychological, and economic risks.
Source: https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

2. Document Writing Policy

In addition to the AI Guidelines for Business, this document has been prepared based on a survey of domestic and international publications and tools.

AISI Japan AI Safety Institute

Domestic and International Publications

Machine Learning Quality Management Guidelines, 4th Edition (National Institute of Advanced Industrial Science and Technology)

The guidelines to classify and organize the quality requirements for AI systems utilizing machine learning in a top-down manner. They provide a structure that enables stakeholders involved in the target system to establish a framework for the objective evaluation of the system's quality.

LLM AI Cyber Security and Governance Checklist (Open Worldwide Application Security Project (OWASP))

It describes the management of cybersecurity risks in the provision and use of AI systems within an organization, including a list of the top 10 critical vulnerabilities in LLM applications.

SP800-115, National Institute of Standards and Technology (NIST)

Comprehensive guidelines for information systems security testing and evaluation.

AI 800-1(Initial Public Draft) (National Institute of Standards and Technology (NIST))

Draft guidelines for risk management of dual-use foundation models.

AI Guidelines for Business(Japan)

Guidelines developed by integrating and updating existing relevant guidelines in Japan in order to respond to rapid technological changes in recent years.

Guide to Red Teaming Methodology on AI Safety

Tools (Organization)

Anthropic

Anthropic has publicly announced that it conducts red teaming from various perspectives and publishes the data sets for red teaming.

Microsoft

Microsoft is conducting red teaming for its services and has published a guide on red teaming.

NVIDIA

Nvidia has conducted red teaming by cross-disciplinary teams within its own company.

OpenAI

OpenAI is recruiting experts to red team their AI model, working to enhance the safety of their products.

Project Moonshot (AI Verify Foundation)

Project Moonshot is an open-source tool developed by the AI Verify Foundation in Singapore, with features to support red teaming exercise.

3. Structure of This Document

Items considered important for conducting red teaming on AI Safety are categorized by type. The table of contents is organized according to the categories to enhance readability.

AISI Japan AI Safety Institute

- The contents of each section of this document are described based on the items organized from a 5W1H perspective.
- The primary target audience is assumed to be AI developers and AI providers. In particular, the target readers are “development and provision managers” and “business executive officers” who are involved in the planning and conducting red teaming.

Type	Examples of items to be described
What (What is red teaming?)	➢ Definition and scope of “red teaming" ➢ AI systems covered in this publication
Why (Why red teaming?)	➢ Purpose of red teaming ➢ Importance and expected effects of red teaming
Who (Who will conduct red teaming?	➢ What roles are the red teaming conductors?
When (When to conduct red teaming?)	➢ Timing of red teaming
Where (where to conduct red teaming?)	➢ Whether it will be performed by your own organization or by a third party
How (How to conduct red teaming?)	➢ How to plan red teaming and what to prepare for it ➢ What threats to assume in red teaming

Intended Audience

AI Developers and AI Providers

Development and Provision Managers

Business Executives Officers

*Readers who are involved in the planning and conducting of red teaming, among those listed on the left.

Main guide to Red Teaming Methodology on AI Safety [Table of Contents]	
1	Introduction
2	About Red Teaming
3	Typical Attack Methods on LLM systems
4	Red Teaming Structure and Roles
5	Timing of Red Teaming and its Process
6	Planning and Preparation
7	Planning and Conducting Attacks
8	Reporting and Developing Improvement Plans
A	Appendix

In the preparation of Version 1.10, Annex (detailed explanation document) and Supplementary document (examples of deliverables) were prepared in addition to main guide. For more details, please refer to page 15 of this document.

4. Scope of Red Teaming

Red teaming in this document describes as an evaluation method to check the effectiveness of response structure and countermeasures for AI Safety in terms of how attackers attack AI systems. The red teaming described here is specifically focused on AI systems(LLM systems) that incorporate large language models.

AISI Japan AI Safety Institute

Scope of red teaming in this document

- **Red teaming** is "**an evaluation method to check the effectiveness of response structure and countermeasures for AI Safety in terms of how attackers attack AI systems**." In this document, red teaming with respect to AI Safety is simply referred to as "red teaming".

- The scope of the red teaming in this document is organized in terms of **(1) Type of AI system** and **(2) Impact of AI system.** (Red teaming is one of the AI Safety evaluation methods, and its scope is consistent with the one described in the "Guide to Evaluation Perspectives on AI Safety.")

Scope of red teaming in this document

The boxes noted in red below indicate the scope of the red teaming in this document.

Type of AI system	**AI systems with LLM as a Component (Contains examples of attack techniques for AI systems that process multimodal inputs, such as images)**	**Impact of AI system**	**The direct impact an AI system can have on the end users**
	General AI systems excluding the above		Impact on people around and society beyond the end users of AI systems

5. Overview of Red Teaming

Red teaming is "an evaluation method to check the effectiveness of response structure and countermeasures for AI Safety in terms of how attackers attack AI systems," and is one of the methods of AI Safety evaluation.

AISI Japan AI Safety Institute

- The purpose of red teaming is to maintain or enhance AI safety by identifying vulnerabilities such as weaknesses and insufficient countermeasures in the target AI system from an attacker's perspective, then mitigating them through system hardening.

Types of Red Teaming

➢ Red teaming can be categorized as follows.

Category of red teaming tests based on prior knowledge of the attack planner/conductor

- **Black-box Test**
  (The attack planner/conductor does not have any prior knowledge of the system, such as its internal structure.)
- **White-box Test**
  (The attack planner/conductor has sufficient knowledge of the system, such as its internal structure.)
- **Gray-Box Test**
  (The attack planner/conductor has partial knowledge of the system, such as its internal structure.)

Category of the environment in which red teaming is conducted

- **Production Environment**
  (Production environment where AI systems are actually put into practice)
- **Staging Environment**
  (Environment for testing and checking for defects in conditions similar to those of the actual production environment)
- **Development Environment**
  (Environment for developing AI systems)

Category of how attack signatures are attempted

- **Red Teaming with Automated Tools**
- **Manual Red Teaming**
- **Red Teaming with AI Agents**

Typical attack methods on LLM systems

➢ Examples of typical attack methods against LLM systems. They should be considered in red teaming.

- **Direct Prompt Injection**
  Attacker directly injects malicious prompts into the AI system
- **Indirect Prompt Injection**
  Attacker indirectly injects malicious prompts into the AI system
- **Prompt Leaking**
  Attacker extracts the designated system prompt
- **Poisoning Attacks**
  Attacker infiltrates manipulated data or model into data or model during training
- **Evasion Attacks**
  Malicious modification of inputs to the AI system to cause unintended behavior
- **Model Extraction Attack**
  An attack to create a model with the same performance as the target system's model by analyzing its inputs and outputs
- **Membership Inference Attacks**
  An attack that identifies whether certain data is included in the training data by analyzing system's inputs and outputs
- **Model Inversion Attacks**
  An attack that recovers information contained in training data by analyzing inputs and outputs

5. Overview of Red Teaming

It is desirable to have diverse stakeholders participate in the conducting of red teaming. It is also desirable to conduct the red teaming not only before the release of the AI system, but also after the start of operation, as needed.

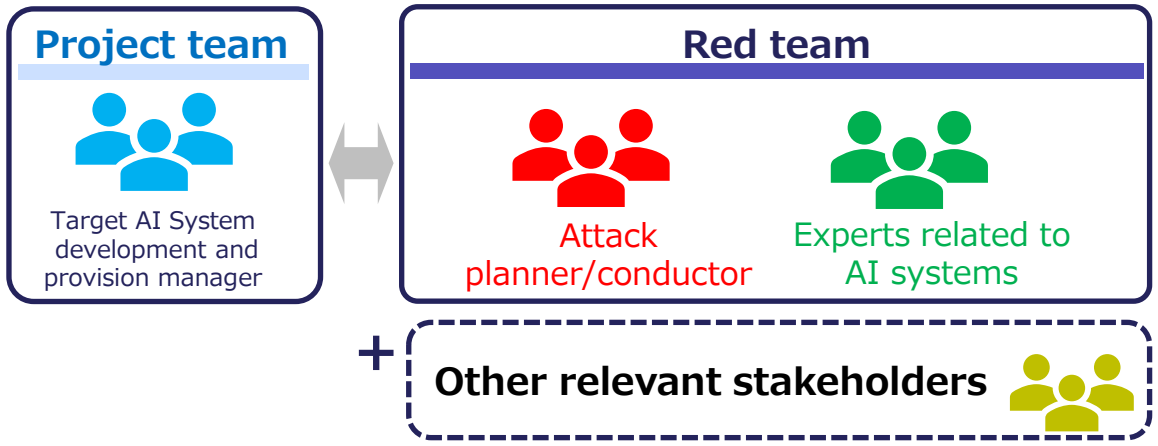
AISI Japan AI Safety Institute

Conducting Structure and Roles

- A red team* should be established in coordination with the project team involved in the development and provision of the AI system subject to red teaming, and a leader or responsible person should be appointed.
- It is basically assumed that the red team includes the "Attack Planner/Conductor," and "Experts related to AI systems."
- Other relevant stakeholders within the organization may be involved as needed.

(*) Team in charge of checking the effectiveness of the response structure and countermeasures for AI Safety in terms of how attackers attack AI systems.

Red team structure



Conducting Timing

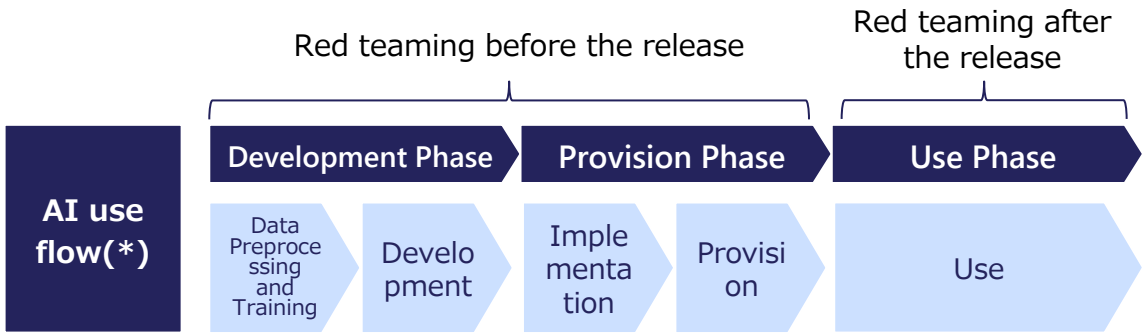
Red teaming before the release

- The first time red teaming is conducted, it should be done before the release of the target AI system.
- However, depending on the scale and complexity of the target AI system, it may be effective to conduct risk analysis in units such as system compornents and system layers, and divide and red teaming at appropriate timing.

Red teaming after the release

- Red teaming is not a one-time task; it should be conducted periodically as needed.

Timing of red teaming in AI use flow



(*)Refer to the AI Guidelines for Business " Correlation between AI business actor and AI use flow".

6. Process of Red Teaming

The red teaming Process consists of three parts: "Planning and Preparation," "Planning and Conducting Attacks," and "Reporting and Developing Improvement Plans."

Process	Items	Chapter in main guide
Process 1: Planning and Preparation	✓ Deciding and launch the red team ✓ Identify and allocate budget and resources, and select and contract third party ✓ Planning ✓ Preparing the environment for red teaming ✓ Confirming escalation flow	Chapter 6.
Process 2: Planning and Conducting Attacks	✓ Developing risk scenarios ✓ Developing attack scenarios ✓ Conducting attack scenarios ✓ Record Keeping during red teaming ✓ After conducting attack scenarios	Chapter 7.
Process 3: Reporting and Developing Improvement Plans	✓ Analyzing the red teaming results ✓ Preparing the report of red teaming results and implementing stakeholder review ✓ Preparing and reporting the final results ✓ Developing and implementing improvement plans ✓ Follow-up after improvement	Chapter 8.

6. Process of Red Teaming（Planning and Preparation）

In “Planning and Preparation,” the project team launches the red team, develop a red teaming plan, and take necessary actions such as allocating budgets.

AISI Japan AI Safety Institute

Process 1: Planning and Preparation

Project team	STEP 1 Deciding and launch the red team	• Provision managers of the target AI system or department of information security and information systems make proposal for the red teaming and make a decision on conducting red teaming. • Red team is established within the organization as described in the proposal.
Project team	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 .
Project team	STEP 2 Identify and allocate budget and resources, and select and contract third party	• 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 prepare "Red Teaming Plan" after reviewing necessary actions such as understanding overview of the target AI system, and collaborate with other relevant stakeholders.
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	STEP 5 Confirming escalation flow	• The red team confirm escalation flow in case of unexpected behavior or failure/trouble due to red teaming conducting.

6. Process of Red Teaming（Planning and Conducting Attacks）

In “Planning and Conducting Attacks,” the red team develops risk and attack scenarios, conduct attack scenarios, keep the records, among other things.

AISI Japan AI Safety Institute

Process 2: Planning and Conducting Attacks

Red team

STEP 6 Developing risk scenarios	• Considering the four factors (system architecture, 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.
STEP 7 Developing attack scenarios	• The attack planner/conductor examines what attacks are actually possible according to the risk scenarios developed, and develop specific attack scenarios to be conducted by red teaming.
STEP 8 Conducting attack scenarios	• Attack scenarios are conducted by dropping specific attack signatures.
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.
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 department of information systems and information security that attacks of red teaming 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.

6. Process of Red Teaming（Reporting and Developing Improvement Plans）

In “Reporting and Developing Improvement Plans,” improvements are made to the items identified as a result of the red teaming. After the results of conducting are reviewed, an improvement plan is developed and implemented, and follow-up on the improvement measures is implemented.

AISI Japan AI Safety Institute

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 primary department in charge of information systems, and the department in charge of information security.
	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 report results of the exercise. • The attack planner/conductor prepares a 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.
Project team	**STEP 13** Preparing and reporting the final results	• The development and provision managers of the target AI system should create a final report of the red teaming, based on the report of red teaming 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 develops an improvement plan, specifying improvement measures to address business risks and other factors. • When formulating improvement measures and developing plans, the project team should determine priorities based on the level of urgency and risk.
	STEP 15 Follow-up after improvement	• The progress of improvement measures implemented based on the improvement plan should be checked at management meetings as appropriate. • After implementing improvements measures, it is advisable to check the configuration 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.

AISI Japan AI Safety Institute

Revision of Guide to Red Teaming Methodology on AI Safety

AI Safety Institute

(March 31, 2025)

Background and Purpose of revision of Guide to Red Teaming Methodology on AI Safety

This document has been revised to accommodate a broader range of red teaming methodologies, making it a more practical guide for reference.

AISI Japan AI Safety Institute

Background

- In September 2024, this document was prepared to provide those involved in the development and provision of AI systems with basic considerations regarding red teaming methodologies to evaluate the countermeasures to risk applied to a target AI system from an attacker's perspective.
- The document is systematically organized based on the "AI Guidelines for Business," as well as a review of literature and relevant stakeholders both domestically and internationally. However, given that Process 2 of red teaming (Planning and Conducting Attacks) requires a high level of expertise, there is a need to enhance the document to serve as a more practical guide.

Purpose

- A study were conducted to expand the scope of evaluation of this document to include multimodal foundation models, and areas for revision were considered.
- Red teaming were conducted in accordance with this document, and revisions were considered to enhance it as a more practical guide. The focus was on use cases involving LLM system using RAG, likely to be adopted by Japanese companies.

*1: AI models that are multimodal (capable of handling multiple modalities) and trained on extensive data, adaptable to a wide range of downstream tasks.

Implementation details for revision of Guide to Red Teaming Methodology on AI Safety

Guide to Red Teaming Methodology on AI Safety as of September 2024 was used as main guide, with some of its contents added, and detailed explanation document as Annex and examples of deliverables as Supplementary document were newly prepared.

AISI Japan AI Safety Institute

The revision points were examined if the evaluation scope is expanded to include multimodal foundation models.

[Revision contents]
Reference information on attack methods concerning multimodal information such as images was added to main guide under “Chapter 3. Typical Attack Methods on LLM Systems."

Based on the results of the implementation aligned with main guide, revisions were considered to make it a practical guide.

[Revision contents]
To help readers better understand and effectively practice each phase of red teaming, the guide as of September 2024 was established as main guide, with some content added, along with Annex that provides detailed explanation document and Supplementary document that provides examples of deliverables.

- ✓ Main guide: flowcharts and other visual aids were included to facilitate a clearer understanding of the red teaming flow for readers.
- ✓ Annex (detailed explanation document): This explains implementation points for conducting red teaming, enabling readers to apply the guide practically.
- ✓ Supplementary document (examples of deliverables) (In Japanese): Examples of deliverables were prepared to help readers visualize the expected deliverables in Japanese.

Add

Prepare new

Guide to Red Teaming Methodology on AI Safety [this document]

Main guide

Main guide [Table of Contents]	
1	Introduction
2	About Red Teaming
3	Typical Attack Methods on LLM systems
4	Red Teaming Structure and Roles
5	Timing of Red Teaming and its Process
6	Planning and Preparation
7	Planning and Conducting Attacks
8	Reporting and Developing Improvement Plans
A	Appendix

Annex (detailed explanation document)

Annex (Detailed Explanation Document) [Table of Contents]	
1	Background and Purpose of the Detailed Explanation Document
2	Role of the Detailed Explanation Document
3	Explanation of Each Process

Supplementary document (examples of deliverables) (In Japanese)

Supplementary document (Examples of Deliverables)	
1	Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (Excel format)
2	The report of red teaming results (PowerPoint format)
3	The final report (PowerPoint format)

Implementation details for revision of Red Teaming Methodology on AI Safety

The Items in each Process of main guide are linked to the content of Annex and Supplementary document, which serves as a more practical reference. In particular, Process 2, which requires a higher level of expertise, is emphasized and explained in detail.

AISI Japan AI Safety Institute

Guide to Red Teaming Methodology on AI Safety

Main guide

This document outlines the fundamental considerations related to red teaming methodology.

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 explains the Items and Implementation points to consider when conducting red teaming in accordance with main guide.

Element	Content	Section
[Overview]	The overview of each Process and Item outlined in the Guide to Red Teaming Methodology on AI Safety are 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.	
[Reference]	An example of the specific Items to be considered in each Process is documented.	

Supplementary document (examples of deliverables)

This provides examples of deliverables that are prepared when conducting red teaming.

Title of Example Deliverables	Content
Developing risk scenarios, attack scenarios, and results of attack scenarios implementation	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	An example of the report of red teaming results, prepared in STEP 12 of main guide, is provided in PowerPoint format.
The final report	An example of the final report, prepared in STEP 13 of main guide, is provided in PowerPoint format.

Implementation details for revision of Guide to Red Teaming Methodology on AI Safety

Red teaming was conducted on an LLM system using RAG, and based on the results, the implementation points and examples of deliverables for each process are documented in Annex and Supplementary document.

AISI Japan AI Safety Institute

[Methods of red teaming]

Conditions of red teaming	Period	Two months from January to February 2025
	Attack vector	Conduct red teaming remotely from company devices through the internal network to the AI system in the development environment, using the input prompt as the attack interface.
	Red teaming scope	Entire AI system
	Definition of attack success	Identification of vulnerabilities that directly cause harm to individuals or organizations, as well as those that may not cause direct harm but have potential concerns for being exploited in other attacks.

Environment of red teaming	Target AI system	Internal business data utilization chatbot service for enterprises using RAG
	User scope	Company employees
	Main provided functions	Employees can upload a complete set of business-related documents to a restricted-access container, enabling the generation of responses based on those documents. This allows employees to use the chatbot as an internal document search service, facilitating flexible information retrieval through a conversational UI.
	AI system configuration	Region: East US 2 Pricing Tier: Standard S0 Base Model: GPT-3.5-Turbo (1106) Rate Limits: 100,000 TPM / 600 RPM *All GPT models in Azure OpenAI Service can be selected and utilized freely. ■ RAG [Search] Azure AI Search (Standard S1) [Data Source] Azure Blob Storage (Standard) [Data Extraction Engine] Azure AI Document Intelligence (Standard S0)

AISI

Japan AI Safety Institute