Guide to Red Teaming Methodology on AI Safety (Version 1.00)

Summary

AI Safety Institute

(September 25, 2024)

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

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.

- 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.

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 Procedure
6	Planning and Preparation
7	Planning and Conducting Attacks
8	Reporting and Developing Improvement Plans
A	Appendix

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 (LLM systems)**	**Impact of AI system**	**The direct impact an AI system can have on the end users**
	Conventional AI systems other than LLM systems		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

- Red teaming aims to discover weaknesses and inadequate countermeasures in the target AI system from the attacker's perspective, and to correct and harden these weaknesses and inadequacies.

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.

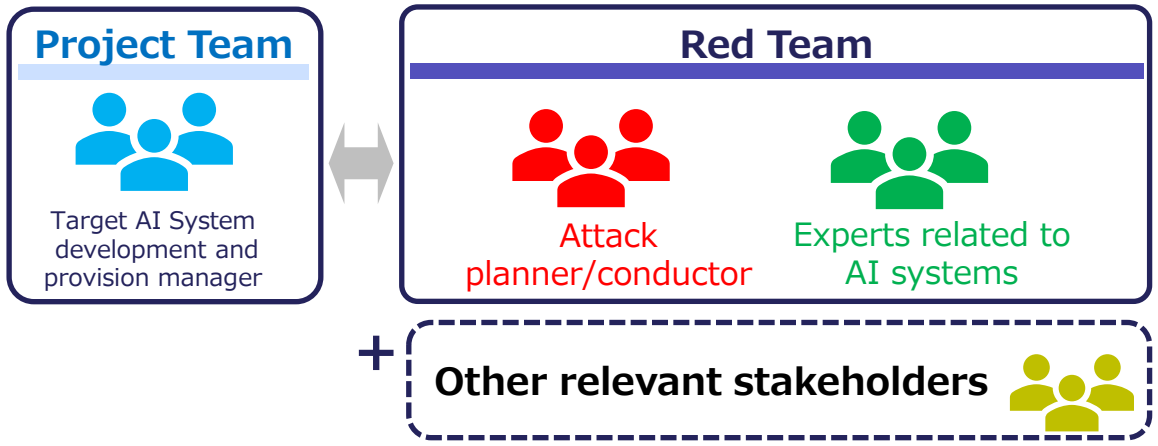
AISI Japan AI Safety Institute

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 will include 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

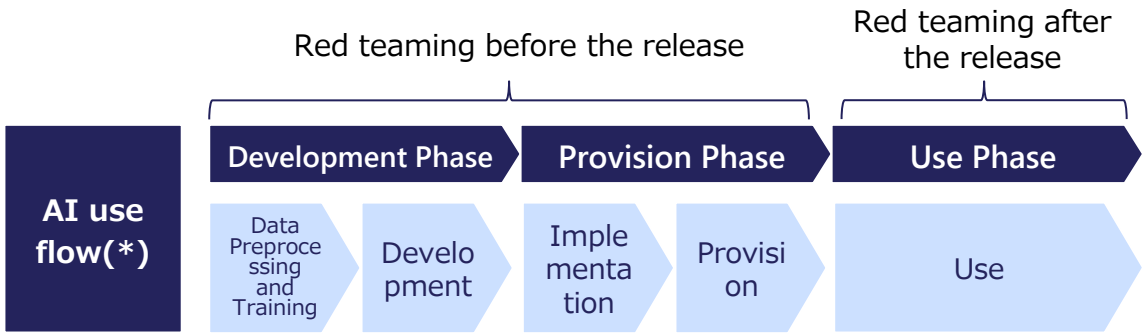
Red teaming before the release

- 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. Procedure of Red Teaming

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

Procedure	Items	Chapter in the document
Procedure 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.
Procedure 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.
Procedure 3: Reporting and Developing Improvement Plans	✓ Analyzing the Red Teaming Results ✓ Preparing the Report of Results and Implementing Stakeholder Review ✓ Preparing and Reporting the Final Results ✓ Developing and Implementing Improvement Plans ✓ Follow-up after Improvement	Chapter 8.

6. Procedure of Red Teaming（Planning and Preparation）

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

AISI Japan AI Safety Institute

Procedure 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 will 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 will 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		• 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 will 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 will be 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 will confirm escalation flow in case of unexpected behavior or failure/trouble due to red teaming conducting.

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

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

AISI Japan AI Safety Institute

Procedure 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 will develop risk scenarios in the target domain and system use cases.
STEP 7 Developing Attack Scenarios	• The attack planner/conductor will examine 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 will notify 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. Procedure 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

Procedure 3: Reporting and Developing Improvement Plans

Red Team	**STEP 11** Analyzing the Red Teaming Results	• The attack planner/conductor will analyze the results obtained from red teaming. • If necessary, additional confirmation of the information to be analyzed will be 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 Results and Implementing Stakeholder Review	• Based on the vulnerabilities discovered during the red teaming exercise, the attack planner/conductor will prepare logs and trails to report results of the exercise. • The attack planner/conductor will prepare a report of 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 red teaming report 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 will develop 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