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

Detailed Explanation Document

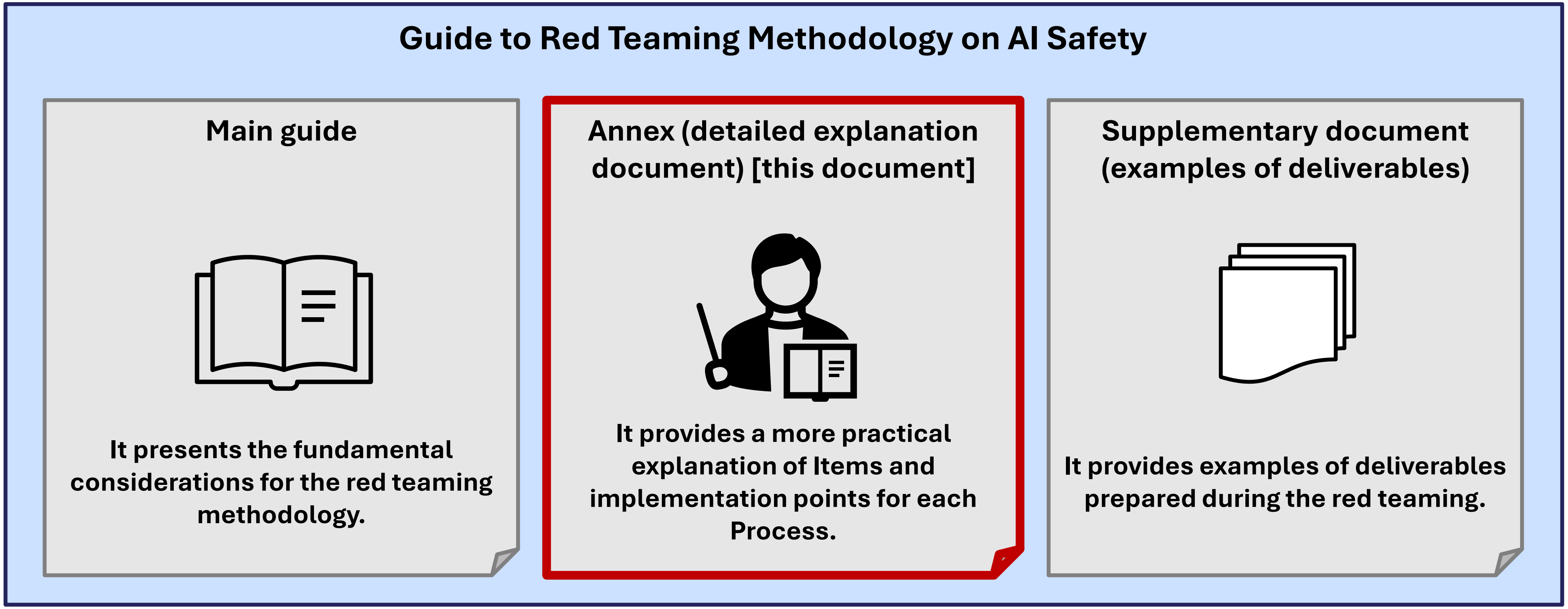
AI Safety Institute
(March 31, 2025)

AISI Japan
AI Safety Institute

Structure of the Guide to Red Teaming Methodology on AI Safety

Guide to Red Teaming Methodology on AI Safety is structured into main guide, Annex (detailed explanation document), and Supplementary document (examples of deliverables).

- Main guide systematically outlines the fundamental considerations for the red teaming methodology, dividing them into three Process (Process 1 to Process 3).
- Annex (detailed explanation document) provides guidance on implementation points when conducting red teaming in accordance with main guide, along with examples of deliverables for each Process.
- Supplementary document (examples of deliverables) presents sample outputs prepared during the red teaming based on main guide, including “Developing risk scenarios, attack scenarios, and results of attack scenarios implementation”, “the report of red teaming results”, and “the final report”.



Guide to Red Teaming Methodology on AI Safety

Main guide

It presents the fundamental considerations for the red teaming methodology.

Annex (detailed explanation document) [this document]

It provides a more practical explanation of Items and implementation points for each Process.

Supplementary document (examples of deliverables)

It provides examples of deliverables prepared during the red teaming.

AISI Japan AI Safety Institute

Table of Contents

1. Background and Purpose of the Detailed Explanation Document

The purpose of this document is to expand it into a more practical resource by conducting red teaming in accordance with Guide to Red Teaming Methodology on AI Safety and presenting insights gained from the results as the detailed explanation document.

Background

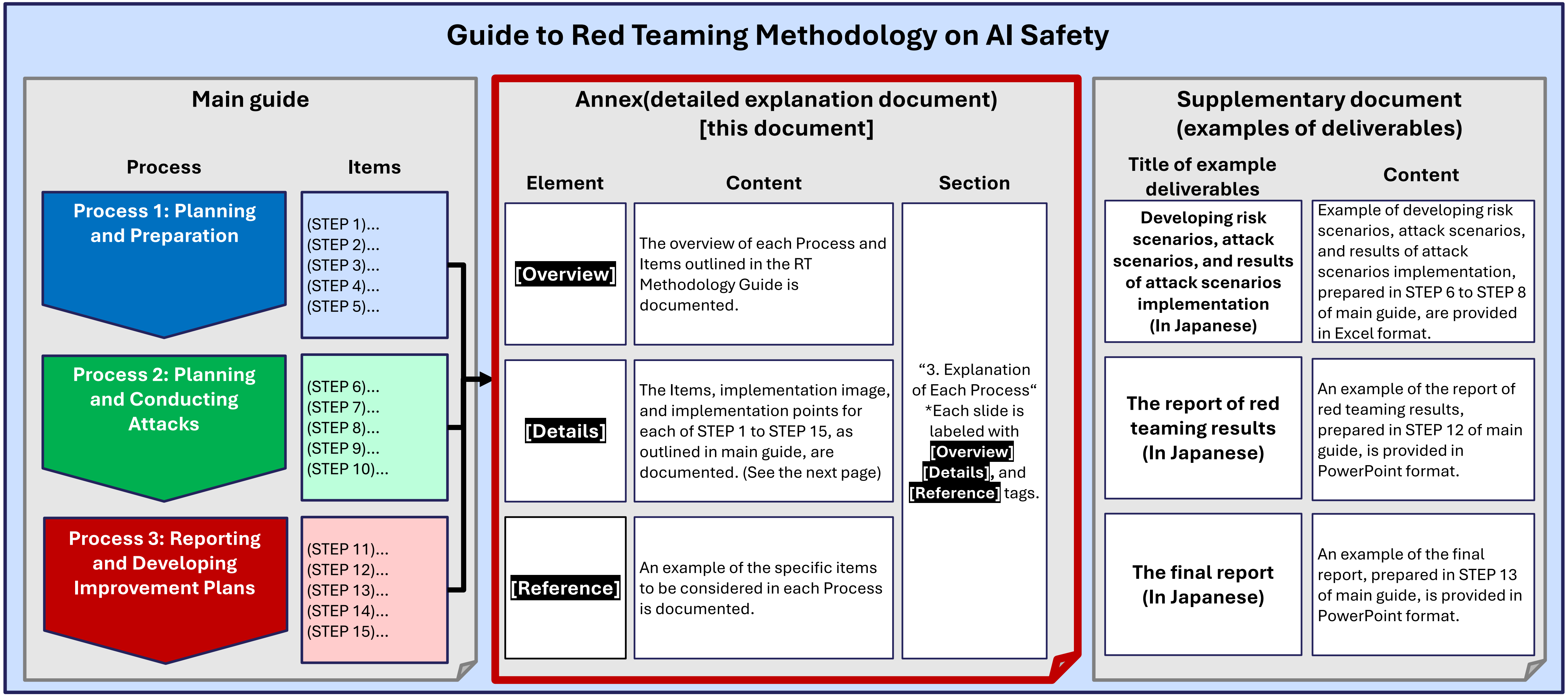
- Red teaming (hereinafter referred to as RT) is **a methodology** used by individuals involved in the development and provision of AI systems **to evaluate the effectiveness of risk mitigation measures applied to the target AI system from an attacker's perspective**.
- In September 2024, the "Guide to Red Teaming Methodology on AI Safety" (hereinafter referred to as RT Methodology Guide) was prepared **to outline the fundamental considerations for RT**.
- The RT Methodology Guide was systematically developed based on the AI Guidelines for Business, as well as a review of domestic and international literature and investigations of relevant industry practices. However, **Process 2 (Planning and Conducting Attacks) in RT requires a high level of expertise, necessitating a more practical guide.** To address this need, RT was conducted on an LLM system utilizing RAG, and insights gained from the results were incorporated into the guide.

Purpose

- The purpose of this document is to expand the RT Methodology Guide into a more practical resource **by conducting RT in accordance with the guide and presenting insights gained from the results in Annex “detailed explanation document” (hereinafter referred to as this document)**.
  *The content presented in this document is merely an example, and organizations may modify and implement it as appropriate to suit their specific needs.

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.



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.

2. Role of the Detailed Explanation Document

In the [Details] section of this document, each of STEP 1 to STEP 15, as outlined in main guide, is explained in terms of RT Items, implementation image, and implementation points.

Items

The Items outlined in main guide are documented.

Implementation image

The implementation image of RT is documented.

Implementation points

When conducting RT in accordance with main guide, key considerations for each STEP are documented as implementation points.

3. Explanation of Each Process

3. Explanation of Each Process

The RT 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 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.

3. Explanation of Each Process

Process 1: Planning and Preparation

3. Explanation of Each Process Process 1

Process 1 | Process 2 | Process 3

1 | 2 | 3 | 4 | 5

AISI Japan AI Safety Institute

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

[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

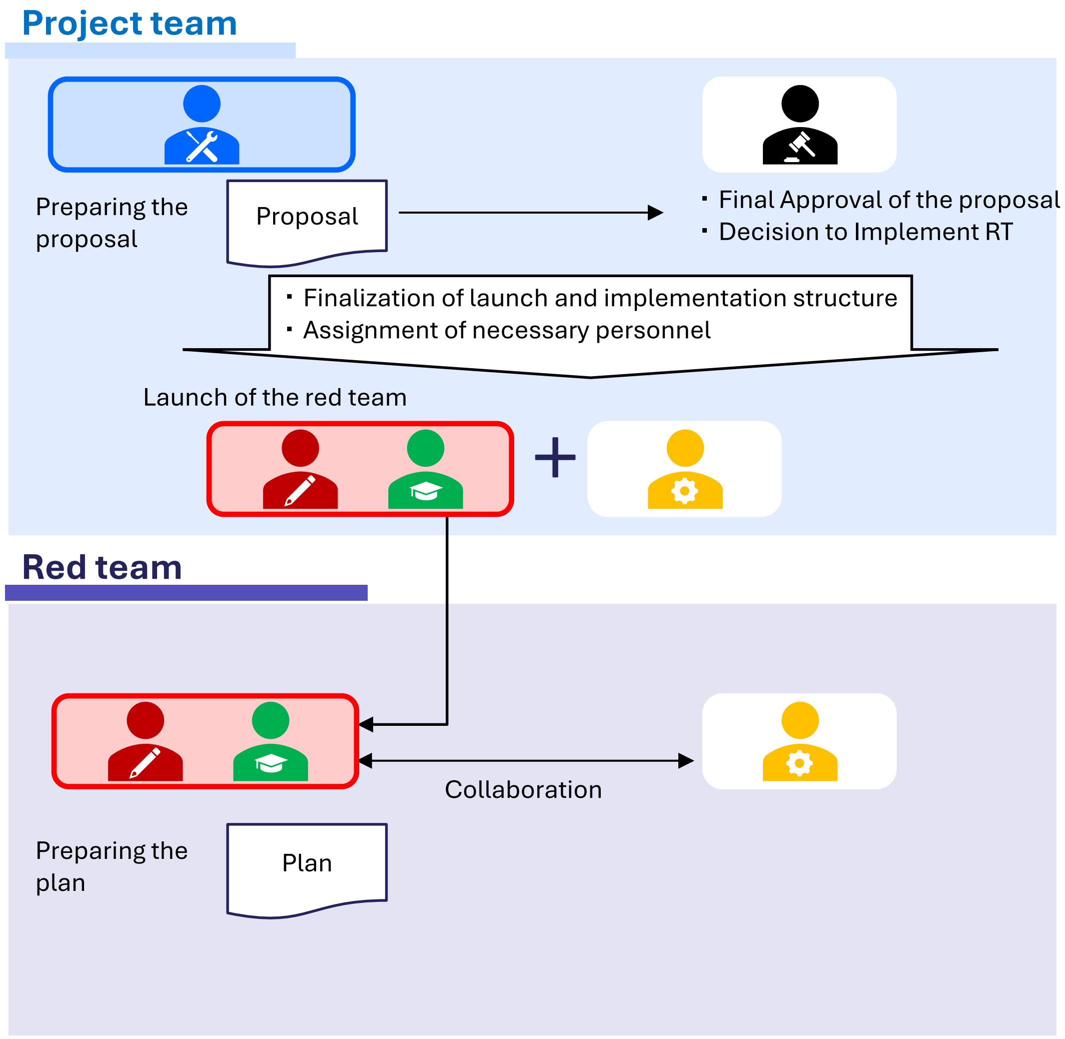
- Provision managers of the target AI system allocate a budget, determine the structure within the organization, and assign the necessary personnel.
- Other resources such as necessary tools are identified and allocated.

- 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.



Project team

Preparing the proposal

Proposal

- Final Approval of the proposal
- Decision to Implement RT

- Finalization of launch and implementation structure
- Assignment of necessary personnel

Launch of the red team

Red team

Collaboration

Preparing the plan

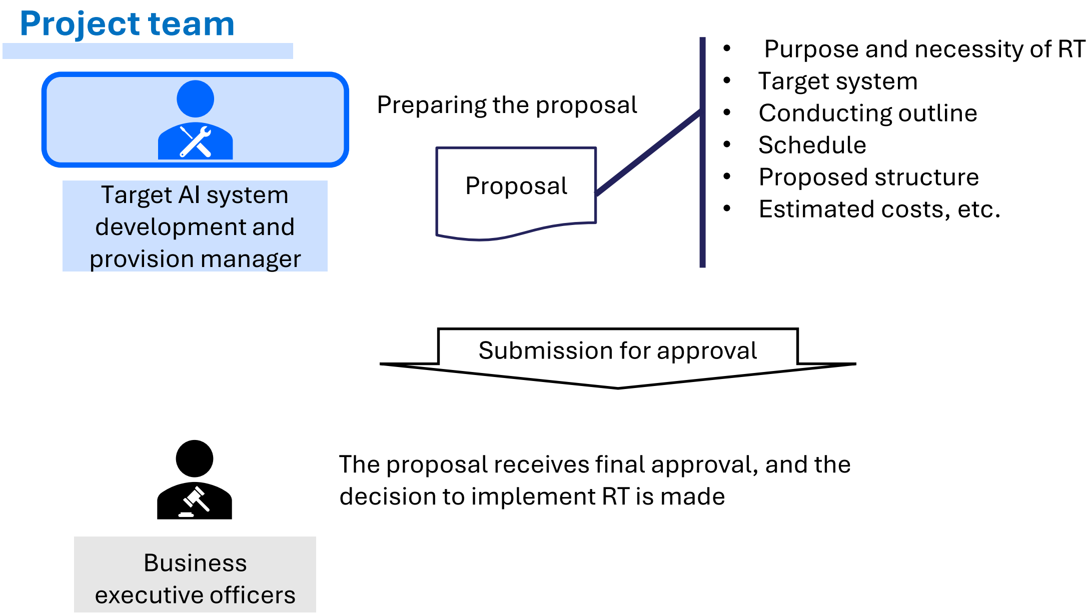
Plan

[Details](STEP 1) Deciding and launch the red team

Items

- Prepare **the proposal** for RT implementation and make the implementation decision.
- Launch the red team within the organization as outlined in the proposal.

Implementation image


Project team
Target AI system development and provision manager
Preparing the proposal
Proposal
Submission for approval
Business executive officers

- Purpose and necessity of RT
- Target system
- Conducting outline
- Schedule
- Proposed structure
- Estimated costs, etc.

The proposal receives final approval, and the decision to implement RT is made

Implementation points

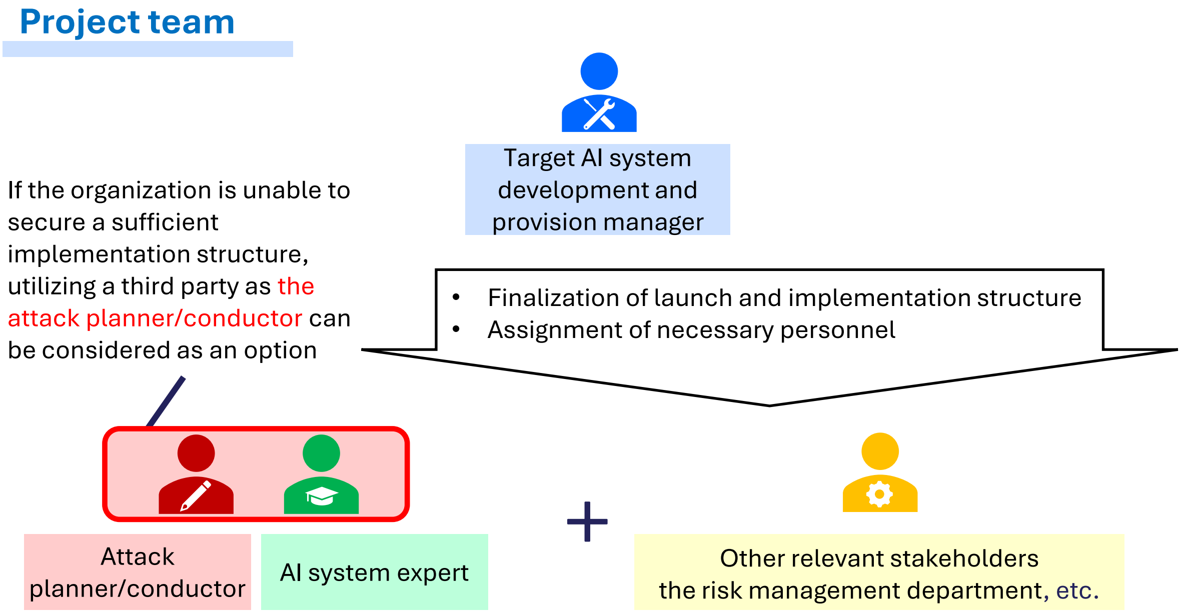
- By gathering detailed information about the department managing the target system for RT, subsequent coordination with stakeholders—such as launching the red team—can be carried out smoothly.
- The composition of the red team should fundamentally include both attack planner/conductor and AI system expert.

[Details](STEP 2) Identify and allocate budget and resources, and select and contract third party

Items

- The target AI system development and provision manager is responsible for securing the budget, finalizing the internal implementation structure, assigning necessary personnel, and ensuring the availability of required tools.

Implementation image


Project team
Target AI system development and provision manager
If the organization is unable to secure a sufficient implementation structure, utilizing a third party as the attack planner/conductor can be considered as an option
• Finalization of launch and implementation structure
• Assignment of necessary personnel
Attack planner/conductor
AI system expert
+
Other relevant stakeholders the risk management department, etc.

Implementation points

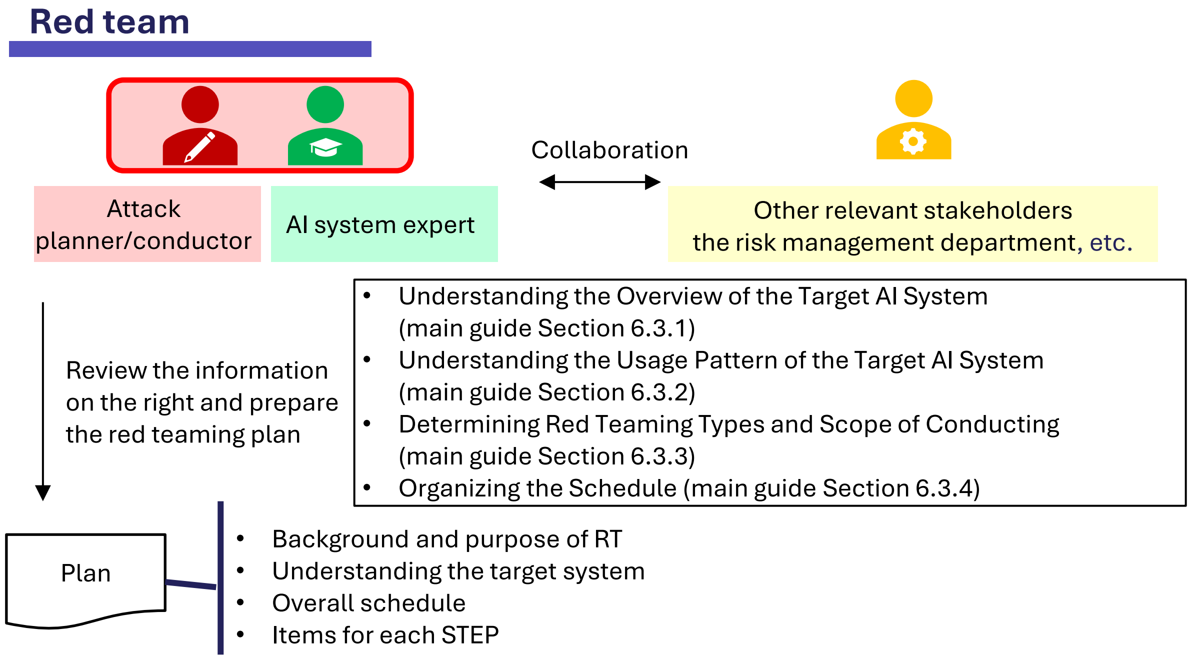
- Since RT implementation requires a high level of expertise, utilizing a third party as an attack planner/conductor can be an option if the internal structure is insufficient. Additionally, as confidential internal information may be handled during the RT process, it is essential to implement robust information security protection measures.

[Details](STEP 3) Planning

Items

- Based on Section 6.3.1 "Understanding the Overview of the Target AI System", Section 6.3.2 "Understanding the Usage Pattern of the Target AI System", and Section 6.3.3 "Determining Red Teaming Types and Scope of Conducting" in main guide, **the necessary items** for RT implementation should be considered. Subsequently, **the red teaming plan** should be prepared, ensuring collaboration with other relevant stakeholders.

Implementation image


Red team
Attack planner/conductor
AI system expert
Collaboration
Other relevant stakeholders
the risk management department, etc.
Review the information on the right and prepare the red teaming plan
Plan

- Understanding the Overview of the Target AI System (main guide Section 6.3.1)
- Understanding the Usage Pattern of the Target AI System (main guide Section 6.3.2)
- Determining Red Teaming Types and Scope of Conducting (main guide Section 6.3.3)
- Organizing the Schedule (main guide Section 6.3.4)

- Background and purpose of RT
- Understanding the target system
- Overall schedule
- Items for each STEP

Implementation points

- As examples of items to be considered in Section 6.3.1 "Understanding the Overview of the Target AI System“ and Section 6.3.2 "Understanding the Usage Pattern of the Target AI System" of main guide, the items listed in [Reference] on P.14 can be considered.
- The red teaming plan should not be limited to the items described in STEP 3, but should be structured with consideration of the entire step from STEP 4 onward.

[Reference](STEP 3) Examples of Items to be considered in Section 6.3.1 and Section 6.3.2

Examples of items to be considered in "Understanding the Overview of the Target AI System"

Category	Item
Understanding the overview of the target AI system	Overall system configuration diagram and network diagram of the AI system
	Use cases of the AI system
	Operational overview of the AI system
	LLM comprising the AI system
	Non-LLM components of the AI system
	Data handled

Examples of items to be considered in "Determining Red Teaming Types and Scope of Conducting"

Category	Item
LLM usage patterns	(A) Usage patterns regarding LLM output
	(B) Usage patterns regarding reference sources of LLM
	(C) Usage patterns regarding LLM itself
Understanding the components other than LLM	Use of commercial plugins and libraries
	Use of OSS plugins and libraries
	Use of proprietary plugins and libraries developed in-house
Existing defense mechanisms	Pre-filtering mechanism to check inputs to the LLM
	Defensive measures in the LLM itself
	Post-filtering mechanism to check outputs from the LLM
	Reinforcement learning with user feedback on inputs and outputs
Other materials to collect	User prompts set for the target LLM
	System prompts
	Deployment environment
	API parameters
	Status of fine-tuning implementation
	Use of user data for training
	Source of training data
	Information on red teaming conducted by other organizations

3. Explanation of Each Process Process 1

Process 1 | Process 2 | Process 3

1 | 2 | 3 | 4 | 5

AISI Japan AI Safety Institute

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

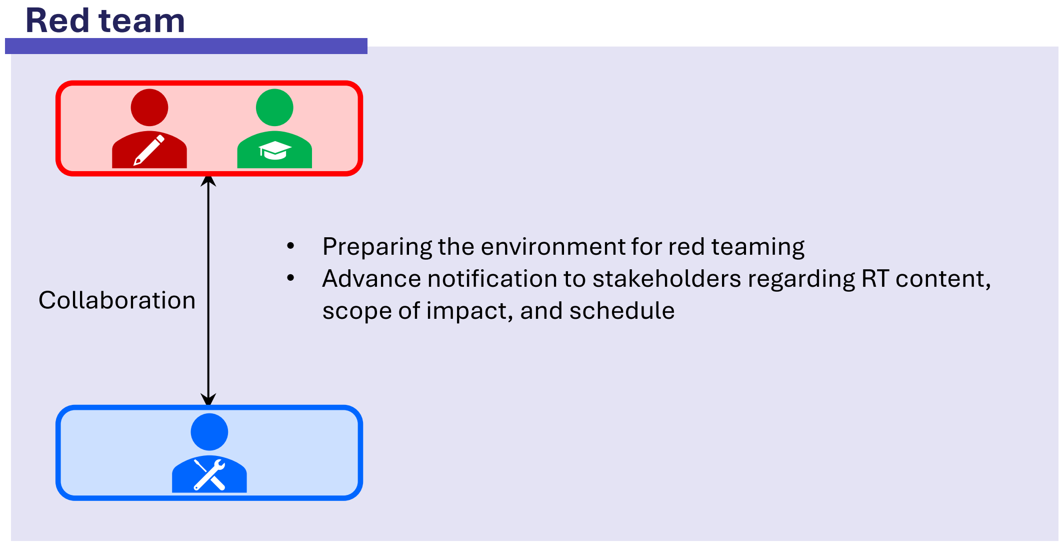
[Legend] ● :Attack planner/conductor ● :AI system expert ● :Target AI system development and provision manager ● :Other relevant stakeholders ● :Business executive officers Project team Red team

Process 1: Planning and Preparation

Red team

STEP 4
Preparing the environment for red teaming

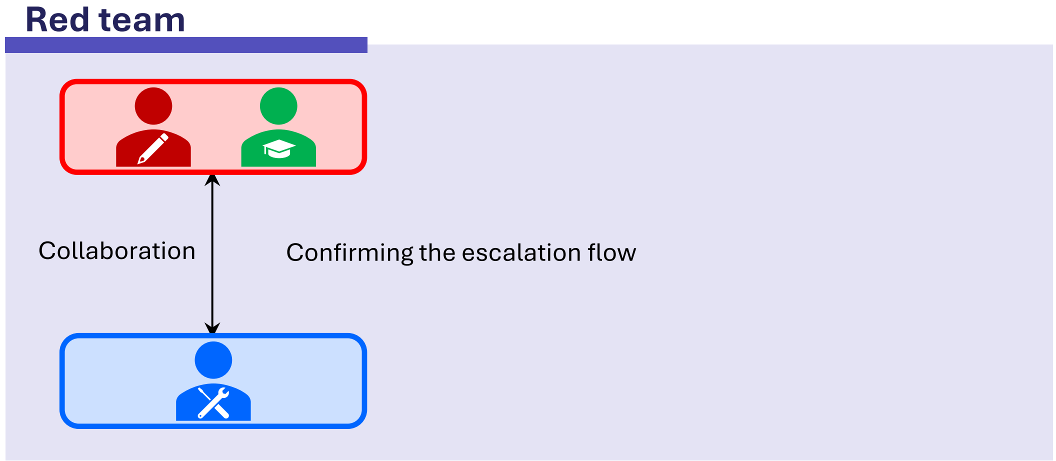
- Prior to conducting red teaming, the content, scope of impact, schedule, and other relevant details should be communicated to stakeholders.
- As needed, stakeholders are informed in advance and requested to temporarily disable monitoring settings, exclude themselves from monitoring, or ignore alerts.



- 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.



Confirming the escalation flow

3. Explanation of Each Process Process 1

Process 1 | Process 2 | Process 3

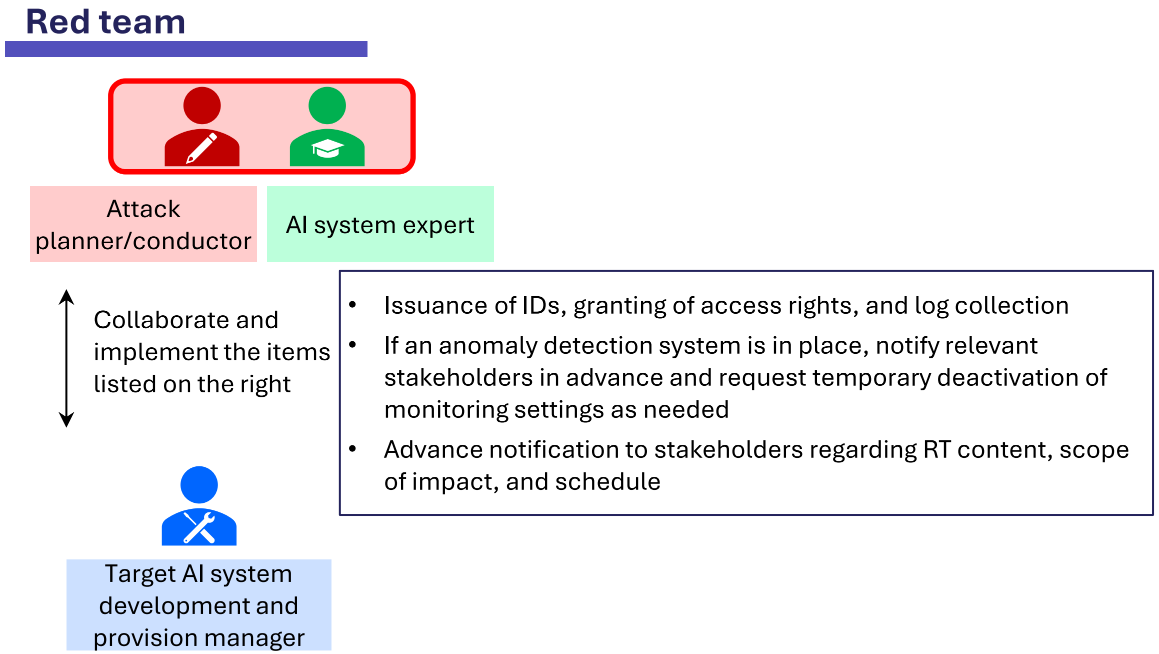
1 | 2 | 3 | 4 | 5

[Details](STEP 4) Preparing the environment for red teaming

Items

- The red team collaborates with the target AI system development and provision manager to make the necessary preparations for the RT execution environment.
- During this process, the content, impact scope, and schedule of the planned RT should be communicated in advance to relevant stakeholders.
- If necessary, stakeholders should be informed beforehand, and requests can be made for temporary deactivation of monitoring settings, exclusion from monitoring targets, or ignoring alerts.

Implementation image


Red team
Attack planner/conductor
AI system expert
Collaborate and implement the items listed on the right
Target AI system development and provision manager

- Issuance of IDs, granting of access rights, and log collection
- If an anomaly detection system is in place, notify relevant stakeholders in advance and request temporary deactivation of monitoring settings as needed
- Advance notification to stakeholders regarding RT content, scope of impact, and schedule

Implementation points

- The execution of RT may result in a large volume of detection logs being generated by the anomaly detection system. Therefore, in addition to preparing the RT execution environment, it is advisable to coordinate in advance with relevant stakeholders—particularly organizations involved with systems that may be affected by the attack scenarios—regarding the RT content, impact scope, and schedule.

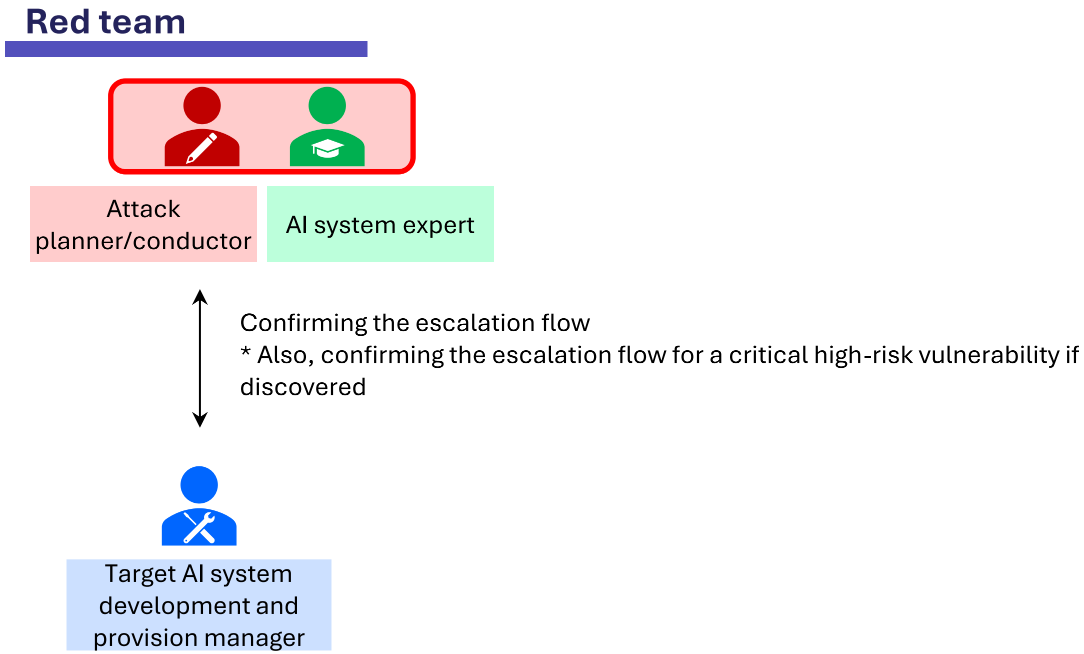
3. Explanation of Each Process Process 1

[Details](STEP 5) Confirming escalation flow

Items

- Before executing RT, the escalation flow should be confirmed to prepare for unexpected behaviors, failures, or issues that may arise.
- If a critical high-risk vulnerability is discovered, the information should be immediately shared with relevant stakeholders without waiting for the completion of the report of red teaming results. In such cases, the escalation flow for urgent reporting should also be confirmed in advance.

Implementation image



Confirming the escalation flow
* Also, confirming the escalation flow for a critical high-risk vulnerability if discovered

Implementation points

- Prior agreements should be established among stakeholders regarding measures for potential failures, operational impacts, or the discovery of critical vulnerabilities. This ensures that any unexpected situations occurring during RT execution can be handled swiftly and effectively.

3. Explanation of Each Process

Process 2: Planning and Conducting Attacks

3. Explanation of Each Process Process 2

AISI Japan AI Safety Institute

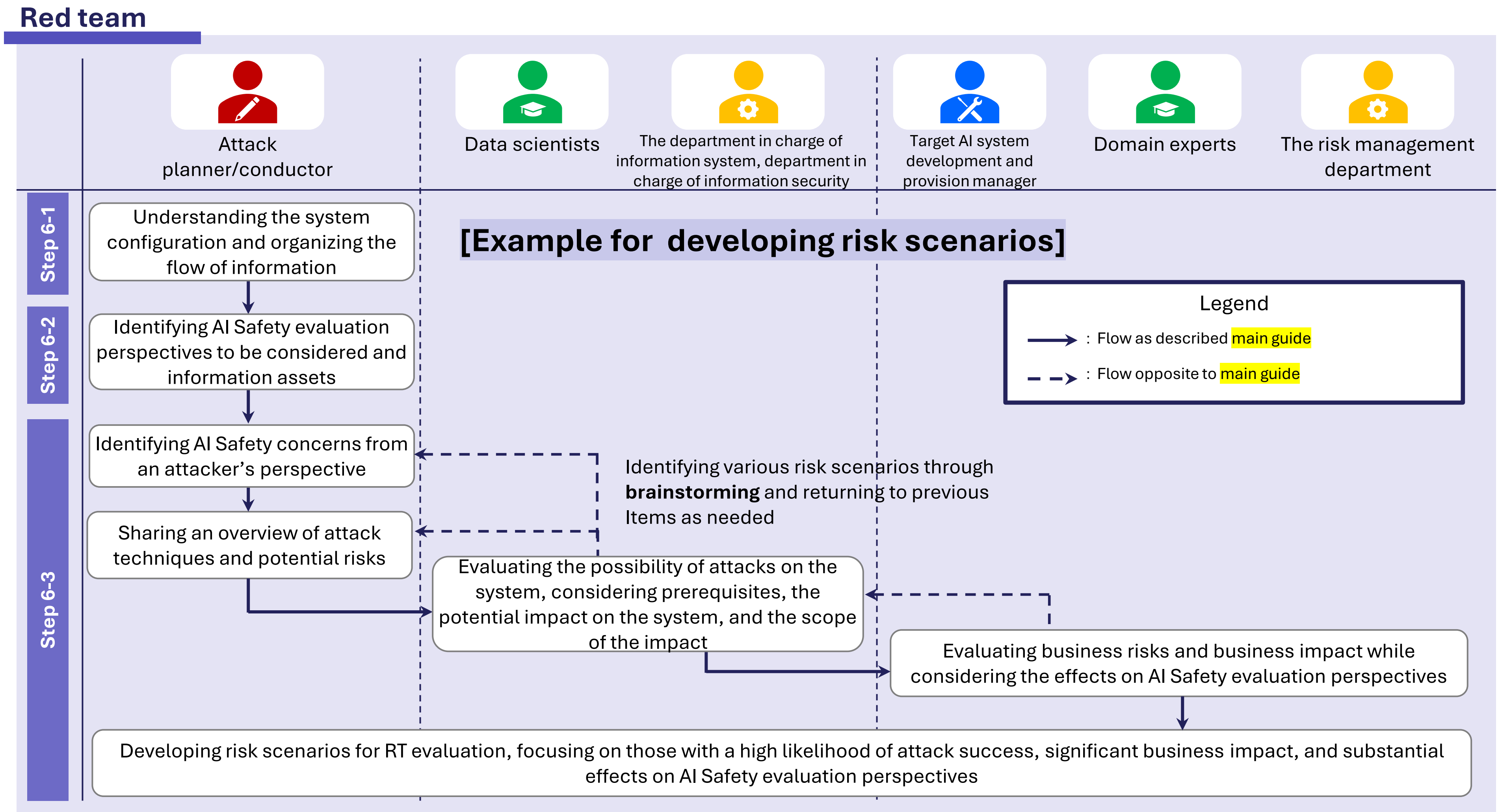
[Overview](STEP 6) Developing risk scenarios

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
Step 6-2
Step 6-3

Understanding the system configuration and organizing the flow of information

Identifying AI Safety evaluation perspectives to be considered and information assets

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

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

[Example for developing risk scenarios]

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

[Details](STEP 6-1) Understanding the system configuration

Items

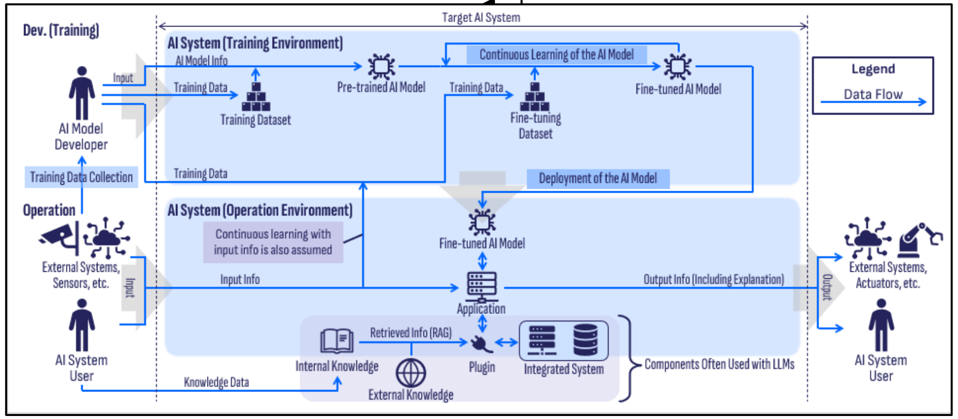
- Based on the information obtained from Section 6.3.1 "Understanding the Overview of the Target AI System" of main guide, **the system configuration** should be identified, **and the flow of information** between LLM inputs/outputs and other components should be organized.

Implementation image

Red team

Attack planner/conductor

Understanding the system configuration and organizing the flow of information between LLM inputs/outputs and other components



Implementation points

- Organize a more detailed flow of information between LLM inputs/outputs and other components based on the system configuration diagram identified in Section 6.3.1 “Understanding the Overview of the Target AI System” of main guide.
- Figure 5 in Section 6.3.1 of main guide provides a reference example of an AI system configuration composed of two environments (development/operation).
- In [Example of deliverables: The report of red teaming results (In Japanese)] *, an example of a system configuration diagram is provided for a RAG-based internal business data utilization chatbot service.

*Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)>The report of red teaming results (In Japanese)

3. Explanation of Each Process Process 2

[Details](STEP 6-2) Identifying AI Safety evaluation perspectives to be considered and information assets to be protected

Items

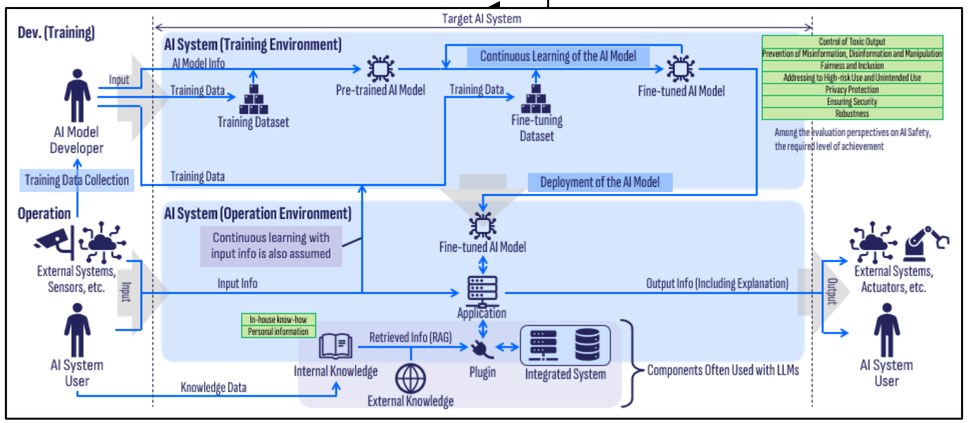
- Identify **the information assets that need protection** based on the services and functions within the system, with a particular focus on critical information that must be safeguarded from attackers.
- Verify **the required levels for each evaluation perspective of AI Safety**.

Implementation image

Red team

Attack planner/conductor

Identifying information assets to be protected from attackers and the required levels for each AI evaluation perspective of AI Safety



Implementation points

- “The required levels for each evaluation perspective of AI Safety” should be defined by each organization, considering the characteristics of the individual AI system and the type of information handled.
- In [Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)] and [Example of deliverables: The report of red teaming results (In Japanese)] *2, an example of the evaluation of required levels for AI Safety perspectives is provided.

*1 Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)*2 Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> The report of red teaming results (In Japanese)

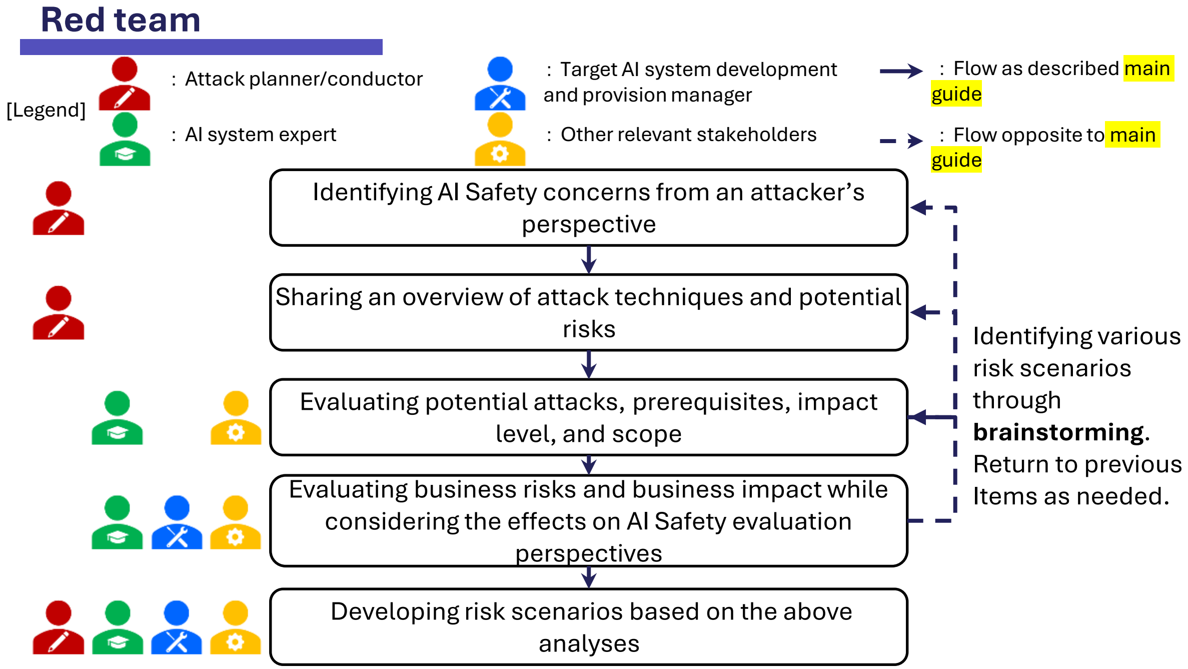
3. Explanation of Each Process Process 2

[Details](STEP 6-3) Developing risk scenarios based on system configuration and usage patterns

Items

- Develop risk scenarios.
- Risk scenarios is a scenario that specifically anticipates potential risks in an AI system and its operational environment, clarifying where threats may arise and their potential impact.
- For example, it can be developed by closely collaborating with relevant stakeholders while following the flowchart shown in the diagram on the right (implementation image).

Implementation image


Red team
[Legend]
: Attack planner/conductor
: AI system expert
: Target AI system development and provision manager
: Other relevant stakeholders
: Flow as described main guide
: Flow opposite to main guide
Identifying AI Safety concerns from an attacker's perspective
Sharing an overview of attack techniques and potential risks
Evaluating potential attacks, prerequisites, impact level, and scope
Evaluating business risks and business impact while considering the effects on AI Safety evaluation perspectives
Developing risk scenarios based on the above analyses
Identifying various risk scenarios through **brainstorming**. Return to previous Items as needed.

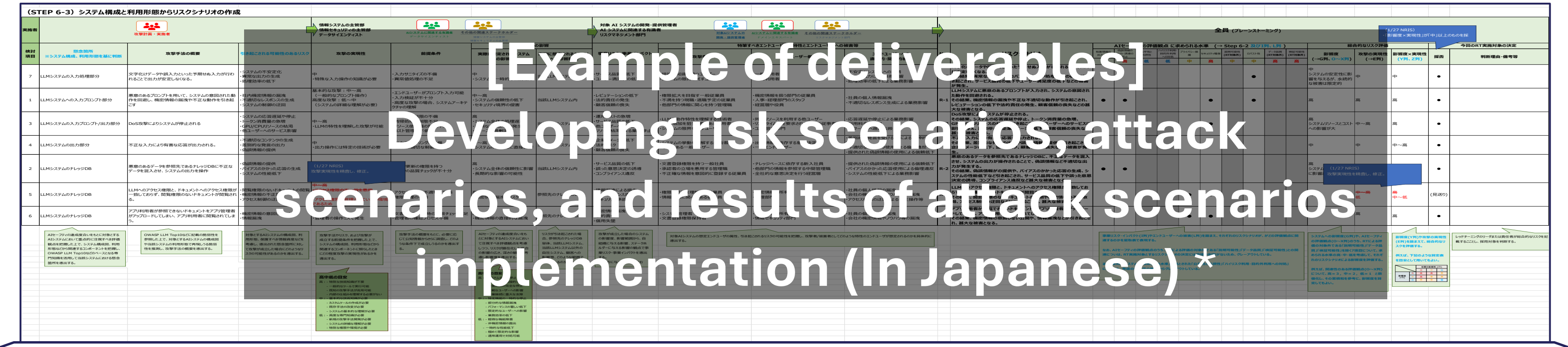
Implementation points

- Consider the items corresponding to the roles of each stakeholder (e.g., attack planner/conductor, the department of information systems and information security) and develop risk scenarios while maintaining close collaboration.
- The [Reference] slide on the next page and [Example of deliverables: Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)] * provide an explanation of the key considerations and procedures for developing risk scenarios.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)

3. Explanation of Each Process Process 2

[Reference](STEP 6-3) Items for consideration by each stakeholder in developing risk scenarios



[Example of deliverables] Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese) *

[Legend] :Attack planner/conductor :AI system expert :Target AI system development and provision manager :Other relevant stakeholders

Actor and Items for consideration in developing risk scenarios											
Actor	Attack planner/conductor			AI system expert (data scientists) Other relevant stakeholders (the department in charge of information system, department in charge of information security)			Target AI system development and provision manager AI system expert (domain experts) Other relevant stakeholders (the risk management department)		All member (brainstorming)		
Row	B	C	D	E	F	G・H	I	J-L	O-X	Y-AA	AB・AC
Items for consideration	Areas of concern	Overview of attack techniques	Potential risks	Feasibility of the attack	Prerequisites	Impact on the system	Business risks and business impact	Characteristics of end user and potential harm	Required levels for AI Safety evaluation perspectives	Overall risk	Determination on RT implementation scope

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)

Note that the format of the example in [Example of deliverables] differs from the Developing Attack Scenarios examples (Tables 4–6) in Section 7.2.3 of main guide. However, the method presented in the deliverables example is only one approach, and organizations may adapt and modify their implementation methods as needed.

3. Explanation of Each Process Process 2

[Reference](STEP 6-3) Example of process for each Item in developing risk scenarios

[Legend] ● :Attack planner/conductor ● :AI system expert ● :Target AI system development and provision manager ● :Other relevant stakeholders

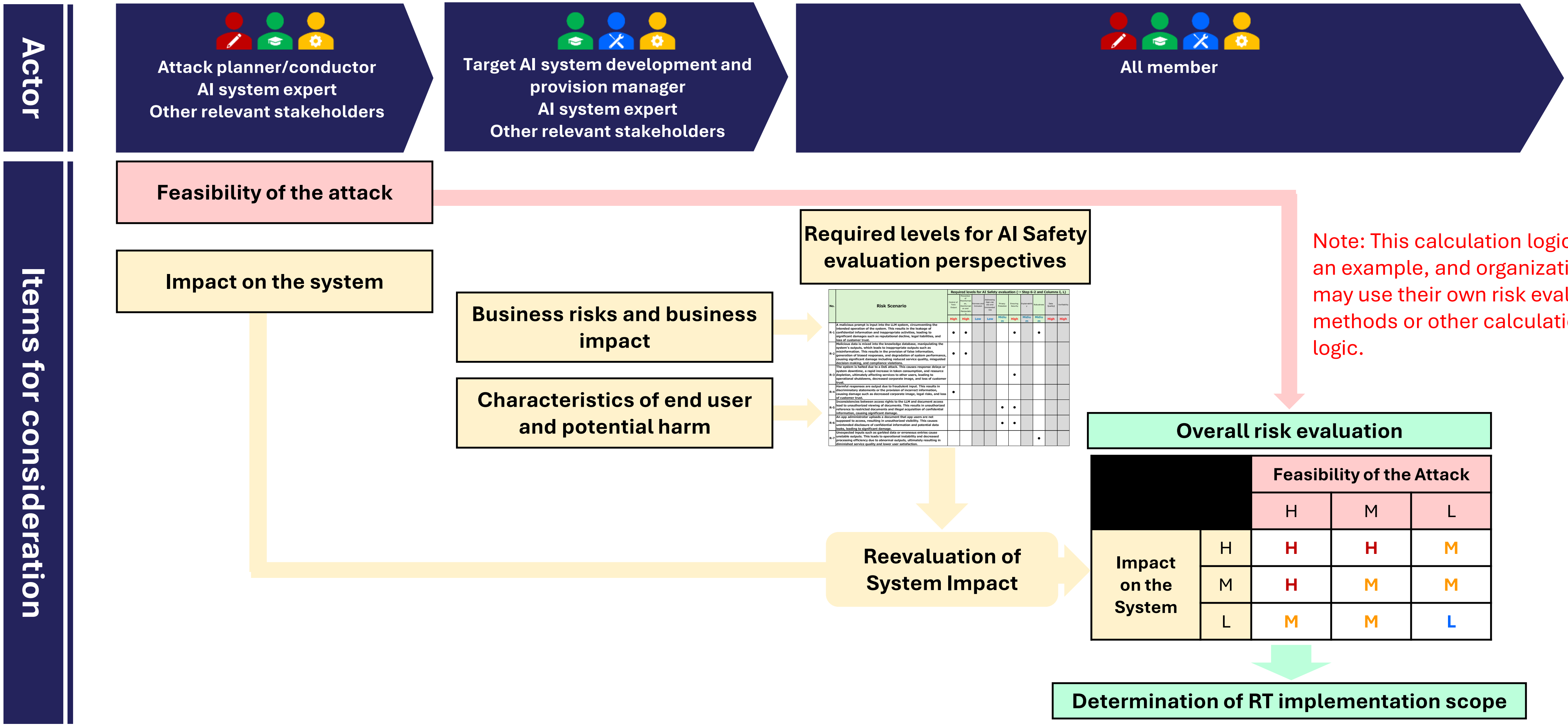
Actor and Items for consideration in developing risk scenarios												
Actor	**Attack planner/conductor**			**AI system expert (data scientists)** **Other relevant stakeholders (the department in charge of information system, department in charge of information security)**				**Target AI system development and provision manager** **AI system expert (domain experts)** **Other relevant stakeholders (the risk management department)**		**All member (brainstorming)**		
Row	**B**	**C**	**D**	**E**	**F**	**G・H**		**I**	**J-L**	**O-X**	**Y-AA**	**AB・AC**
Items for consideration	**Areas of concern**	**Overview of attack techniques**	**Potential risks**	**Feasibility of the attack**	**Prerequisites**	**Impact on the system**		**Business risks and business impact**	**Characteristics of end user and potential harm**	**Required levels for AI Safety evaluation perspectives**	**Overall risk**	**Determination of RT implementation scope**
Example of process	Derive potential attack targets from the system configuration diagram of the target AI system by identifying areas of concern. Based on the required levels for AI Safety evaluation perspectives, determine key evaluation perspectives in advance and conduct the analysis utilizing expert knowledge.	Identify potential vulnerabilities for the listed areas of concern and derive an overview of possible attack techniques. Refer to known vulnerability lists such as OWASP LLM Top 10, and analyze the system configuration and usage patterns of the target AI system to evaluate relevant attack methods.	Consider the system configuration, usage patterns, and critical information assets of the target AI system, identify the potential risks that may arise if an attack succeeds against the identified areas of concern.	Evaluate the feasibility of the identified areas of concern and attack techniques by determining how likely the attacks can be executed. Analyze the attack methods, associated risks, and prerequisites for a successful attack, while considering the system configuration and usage patterns of the target AI system.	Based on the overview of attack techniques, conduct additional research if necessary to determine the conditions under which the attack could successfully be executed.	■Expected level of Impact on the System Determine the extent of the system impact if the identified risk materializes. Evaluate this based on the required levels for AI Safety evaluation perspectives, considering the key evaluation perspectives that should be prioritized.	■Expected scope of Impact on the System Determine the scope of impact if the risk materializes, including the referenced knowledge database alone, the LLM system, internal systems other than the LLM system, and effects on customers.	Derive business risks and business impact by evaluating the level and scope of impact on the system if an attack succeeds, considering both the effects on the organization itself and stakeholders. *Include the impact on both parties.	Identify the characteristics of the target system's expected end user, considering both attackers and victims. For attackers, list their skills and motivations to use as reference information for developing risk scenarios. For victims, outline end user attributes and enumerate the direct and indirect harm that could result.	Determine which evaluation perspectives the risk scenarios aligns with and create matrix table. Refer to the required levels for AI Safety evaluation perspectives defined in the reference materials (*) to ensure comprehensive coverage. If any evaluation perspectives are insufficiently addressed, the risk scenarios are redeveloped.	The RT leader or responsible person derives the overall risk based on the feasibility of the attack and its impact on the system. Refer to the next page for details on the calculation method and relationships of overall risk.	The RT leader or responsible person determines whether to proceed with RT implementation based on the overall risk evaluation results and documents the decision along with the rationale.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)> "(STEP 6-2) The required levels for each evaluation perspective of AI Safety" sheet

[Reference](STEP 6-3) Example of the overall risk evaluation process

- As an example of an overall risk evaluation method, the feasibility of the attack and its impact on the system can be considered.
- The impact on the system should be assessed by considering business risks and impacts, the characteristics of end users, and the required levels for AI Safety evaluation perspectives.

[Legend] :Attack planner/conductor :AI system expert :Target AI system development and provision manager :Other relevant stakeholders


Actor
Attack planner/conductor
AI system expert
Other relevant stakeholders
Target AI system development and provision manager
AI system expert
Other relevant stakeholders
All member
Items for consideration
Feasibility of the attack
Impact on the system
Business risks and business impact
Characteristics of end user and potential harm
Required levels for AI Safety evaluation perspectives
Reevaluation of System Impact
Overall risk evaluation
Determination of RT implementation scope

No.	Risk Scenario	Required levels for AI Safety evaluation (→ Step 6-2 and Columns I, L)									
		Control of Toxic Output	Prevention of Misinformation, Disinformation and Manipulation	Fairness and Inclusion	Addressing High-risk Use and Unintended Use	Privacy Protection	Ensuring Security	Explainability	Robustness	Data Quality)	Verifiability
		High	High	Low	Low	Midium	High	Midium	Midium	High	High
R-1	A malicious prompt is input into the LLM system, circumventing the intended operation of the system. This results in the leakage of confidential information and inappropriate activities, leading to significant damages such as reputational decline, legal liabilities, and loss of customer trust.	●	●				●		●		
R-2	Malicious data is mixed into the knowledge database, manipulating the system's outputs, which leads to inappropriate outputs such as misinformation. This results in the provision of false information, generation of biased responses, and degradation of system performance, causing significant damage including reduced service quality, misguided decision-making, and compliance violations.	●	●								
R-3	The system is halted due to a DoS attack. This causes response delays or system downtime, a rapid increase in token consumption, and resource depletion, ultimately affecting services to other users, leading to operational shutdowns, decreased corporate image, and loss of customer trust.						●				
R-4	Harmful responses are output due to fraudulent input. This results in discriminatory statements or the provision of incorrect information, causing damage such as decreased corporate image, legal risks, and loss of customer trust.	●									
R-5	Inconsistencies between access rights to the LLM and document access lead to unauthorized viewing of documents. This results in unauthorized reference to restricted documents and illegal acquisition of confidential information, causing significant damage.					●	●				
R-6	An app administrator uploads a document that app users are not supposed to access, resulting in unauthorized visibility. This causes unintended disclosure of confidential information and potential data leaks, leading to significant damage.					●	●				
R-7	Unexpected inputs such as garbled data or erroneous entries cause unstable outputs. This leads to operational instability and decreased processing efficiency due to abnormal outputs, ultimately resulting in diminished service quality and lower user satisfaction.								●		

		Feasibility of the Attack		
		H	M	L
Impact on the System	H	H	H	M
	M	H	M	M
	L	M	M	L

Note: This calculation logic is only an example, and organizations may use their own risk evaluation methods or other calculation logic.

[Reference](STEP 6-3) Example of metrics for the feasibility of attack, impact on the system, and risk evaluation

- Classifying the feasibility of an attack and its impact on the system can help derive overall risk evaluation metrics.

Levels (Scoring)	Feasibility of the attack	Impact on the system
Explanation	Determine the likelihood of attack execution (occurrence). The following are examples for each level.	Determine the impact on the system. The following are examples for each level.
H(+3)	-No specialized technical knowledge required -Executable with common tools -Known attack techniques can be reused -No need to understand internal mechanisms	-Complete system shutdown -Direct leakage of confidential information -Total failure of critical functions -Widespread impact on users -Severe disruption to business continuity
M(+2)	-Basic technical knowledge required -Creation of custom tools needed -Modification of existing attack techniques required -Basic understanding of the system necessary	-Temporary shutdown of specific functions -Partial information leakage -Significant performance degradation -Limited impact on specific users
L(+1)	-Advanced expertise required -Development of new attack techniques needed -Detailed understanding of the system necessary -Special privileges or specific environments required	-Minor functional disruptions -Exposure of non-confidential information -Temporary performance degradation -Extremely limited impact -Manageable within normal operations

Overall risk evaluation

		Feasibility of the attack		
		H(+3)	M(+2)	L(+1)
Impact on the system	H(+3)	H(6)	H(5)	M(4)
	M(+2)	H(5)	M(4)	M(3)
	L(+1)	M(4)	M(3)	L(2)

[Criteria]

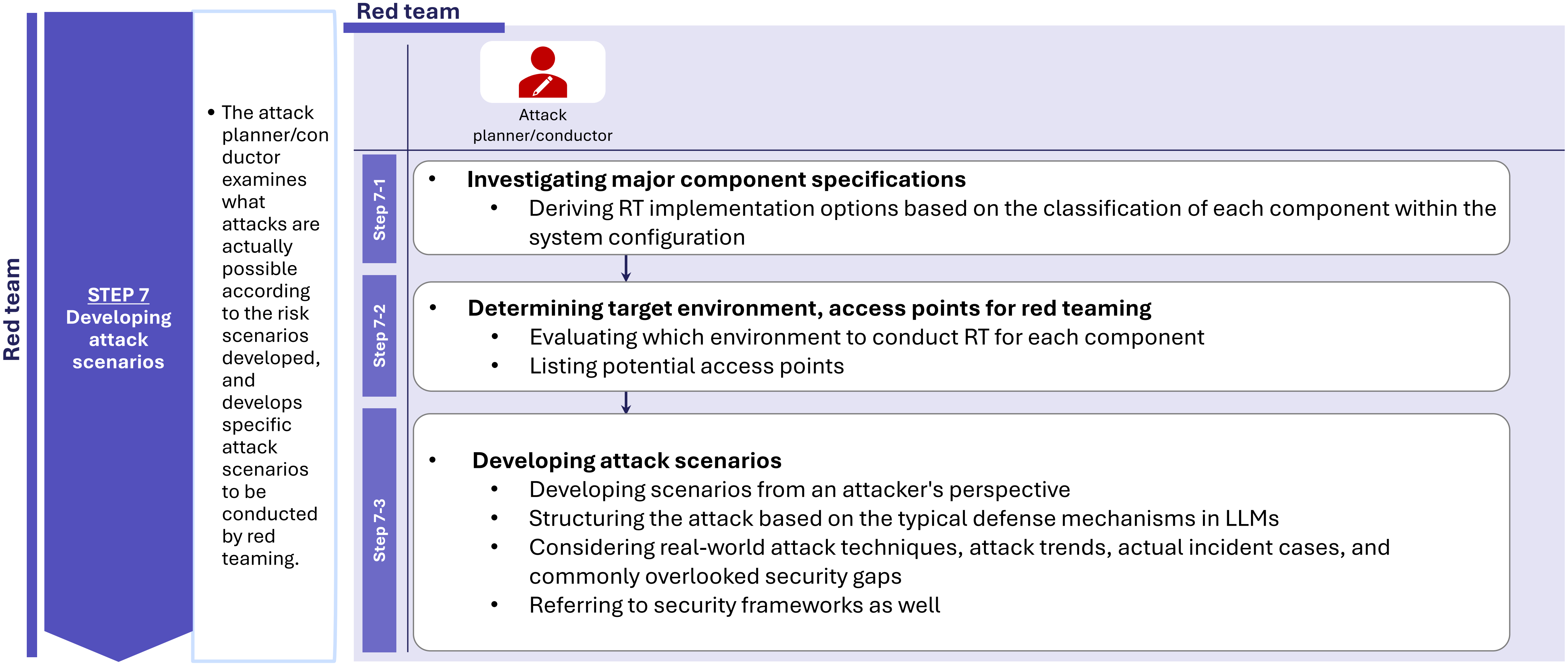
H	Requires action in the current system
M	Requires action in the current system or during the next replacement
L	No action required in the current system

Note: This calculation logic is only an example, and organizations may use their own risk evaluation methods or other calculation logic.

3. Explanation of Each Process Process 2

[Overview](STEP 7) Developing attack scenarios

Process 2: Planning and Conducting Attacks



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.

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

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

- **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

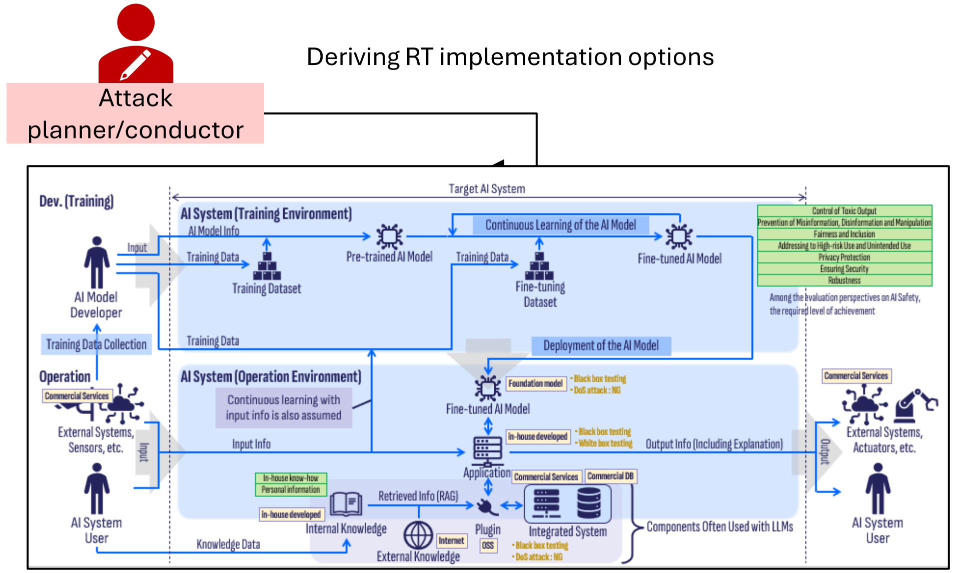
[Details](STEP 7-1) Investigating major component specifications

Items

- Based on the information obtained from Section 7.1.1 "Understanding the System Configuration" and Section 6.3.3 "Determining Red Teaming Types and Scope of Conducting" in main guide, classify each component within the system configuration as a commercial service, open-source software (OSS), or in-house development. Using the classification results, derive detailed options for RT implementation methods.

Implementation image

Red team



Attack planner/conductor

Deriving RT implementation options

Implementation points

- By investigating major component specifications, it becomes clear whether white-box testing is feasible or if only black-box testing is feasible. This understanding allows for the derivation of detailed options for RT implementation methods.

3. Explanation of Each Process Process 2

Japan AI Safety Institute

[Details](STEP 7-2) Determining target environment, access points for red teaming

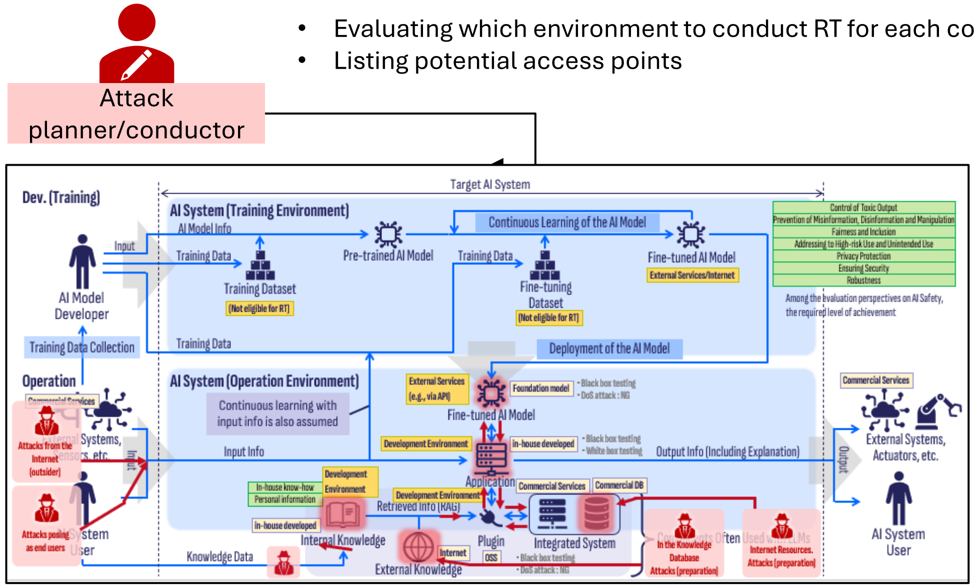
Items

- Based on the information obtained from Section 6.3.3 "Determining Red Teaming Types and Scope of Conducting" in main guide, evaluate the environment in which RT is conducted for each component. Additionally, consider the access points for RT execution.

Implementation image

Red team

- Evaluating which environment to conduct RT for each component
- Listing potential access points



Implementation points

- For each component, evaluate whether RT should be conducted in the in-operation environment, staging environment, or development environment.
- While considering access points for RT execution, this step should be limited to identifying possible access point options.

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

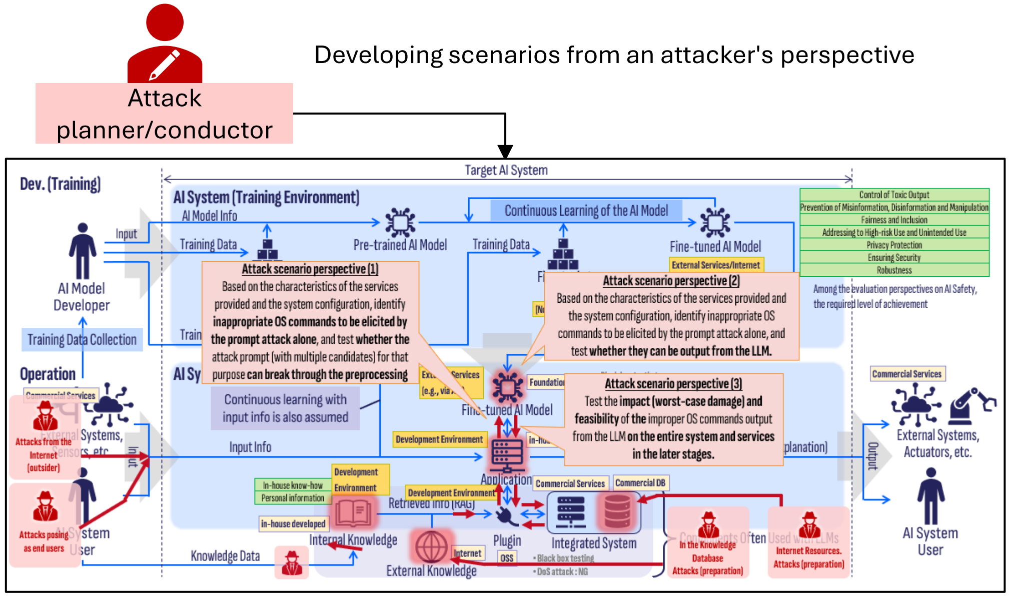
[Details](STEP 7-3) Developing attack scenarios

Items

- The attack planner/conductor develops attack scenarios based on the risk scenarios.
- Attack scenarios are a plan that defines, from an attacker’s perspective, which environment to target, which access points to use, and what combination of attack techniques to employ, based on a specific risk scenario.
- It is effective to consider attack scenarios from multiple perspectives.

Implementation image

Red team



Attack scenario perspective (1)
Based on the characteristics of the services provided and the system configuration, identify **inappropriate OS commands to be elicited by the prompt attack alone**, and test **whether the** attack prompt (with multiple candidates) for that purpose **can break through the preprocessing**

Attack scenario perspective (2)
Based on the characteristics of the services provided and the system configuration, identify inappropriate OS commands to be elicited by the prompt attack alone, and test **whether they can be output from the LLM.**

Attack scenario perspective (3)
Test the **impact (worst-case damage) and feasibility** of **the** improper OS commands output from the LLM **on the entire system and services in the later stages.**

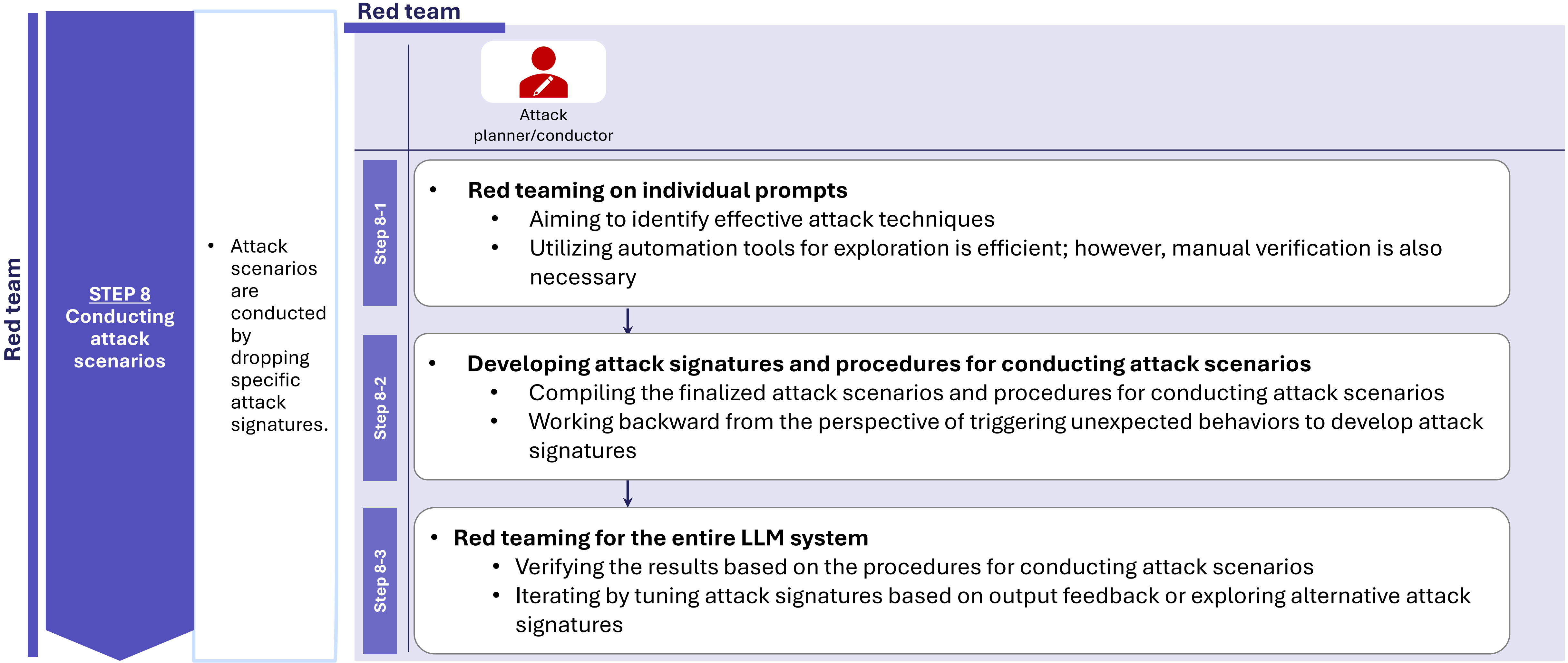
Implementation points

- Considering representative defense mechanisms in LLM systems, reported attack techniques, attack trends, real-world incidents, and commonly overlooked security measures can lead to the effective execution of RT.
- An example of developing attack scenarios is provided in [Example of deliverables: Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)] *.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)

[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

3. Explanation of Each Process Process 2

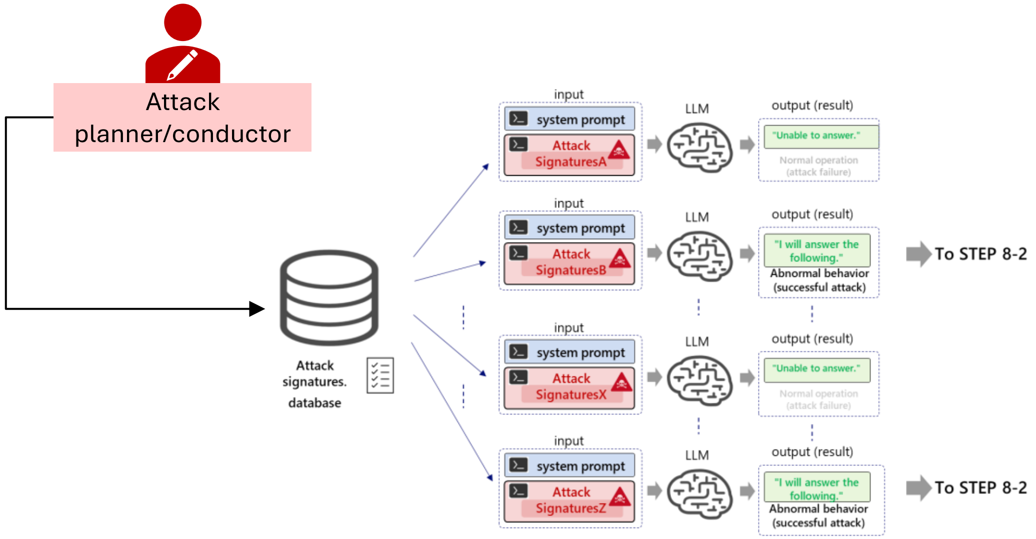
[Details](STEP 8-1) Red teaming on individual prompts

Items

- To identify fundamental vulnerabilities and attack techniques in LLMs, conduct RT targeting individual prompts.
- Input attack signatures into the target system.
- An attack signature refers to a specific input or pattern used to execute a particular attack technique.

Implementation image

Red team



Implementation points

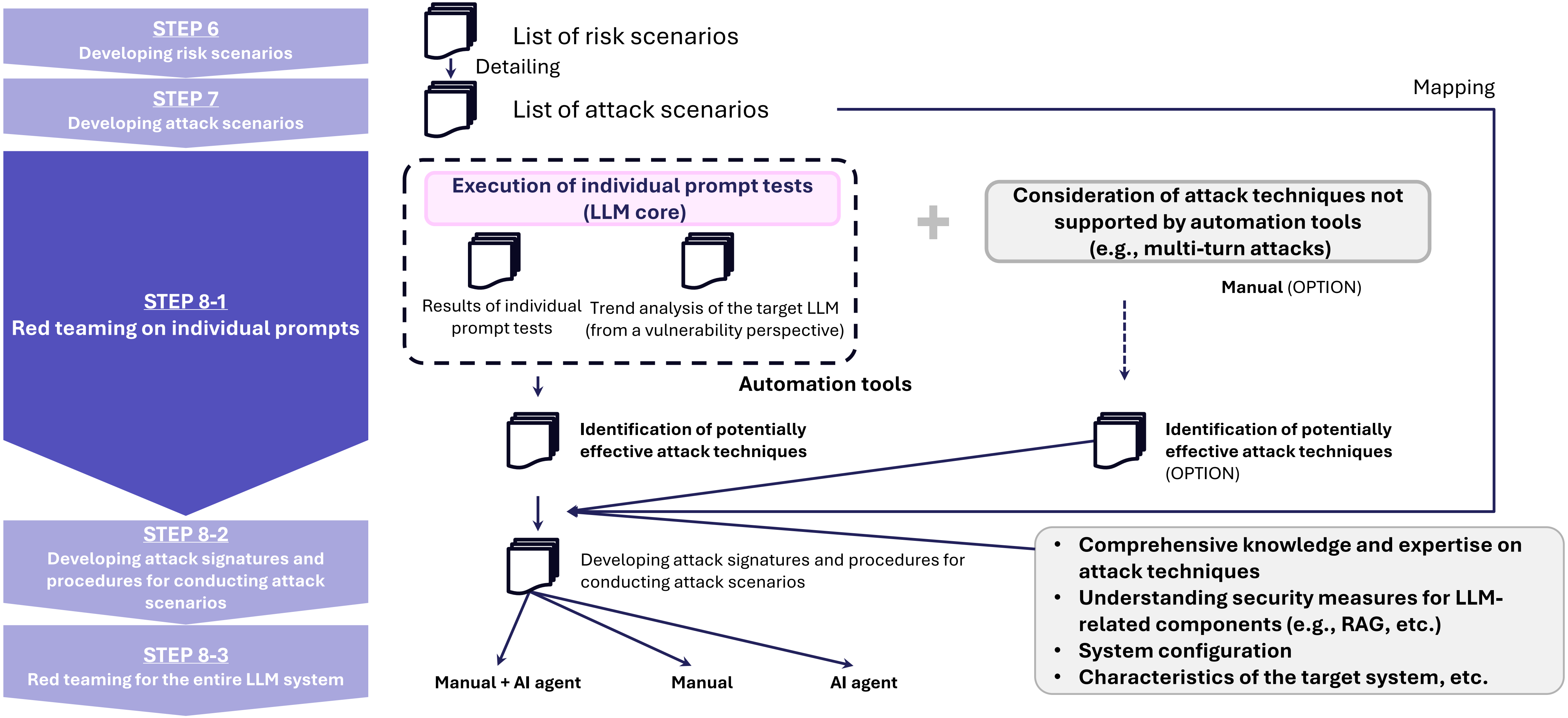
- In RT targeting individual prompts, a large number of attack signatures that can be prepared independently of the target system are input to identify effective attacks. The next page explains the positioning of RT for individual prompts.
- The results of RT for individual prompts using automation tools are provided in [Example of deliverables:Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)】 *.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)

3. Explanation of Each Process Process 2

[Reference]Role of red teaming on individual prompts

- During RT for individual prompts, a large number of attack signatures that can be prepared independently of the target system are fed into identify effective attacks.
- Since prompt injection attacks need to be conducted at scale, leveraging automation tools is recommended. However, for attack techniques not yet supported by these tools, such as multi-turn attacks, manual execution should also be considered.



STEP 6
Developing risk scenarios

STEP 7
Developing attack scenarios

STEP 8-1
Red teaming on individual prompts

STEP 8-2
Developing attack signatures and procedures for conducting attack scenarios

STEP 8-3
Red teaming for the entire LLM system

List of risk scenarios

Detailing

List of attack scenarios

Mapping

Execution of individual prompt tests (LLM core)

Results of individual prompt tests

Trend analysis of the target LLM (from a vulnerability perspective)

Automation tools

Consideration of attack techniques not supported by automation tools (e.g., multi-turn attacks)

Manual (OPTION)

Identification of potentially effective attack techniques

Identification of potentially effective attack techniques (OPTION)

Developing attack signatures and procedures for conducting attack scenarios

- Comprehensive knowledge and expertise on attack techniques
- Understanding security measures for LLM-related components (e.g., RAG, etc.)
- System configuration
- Characteristics of the target system, etc.

Manual + AI agent

Manual

AI agent

[Details](STEP 8-2) Developing attack signatures and procedures for conducting attack scenarios

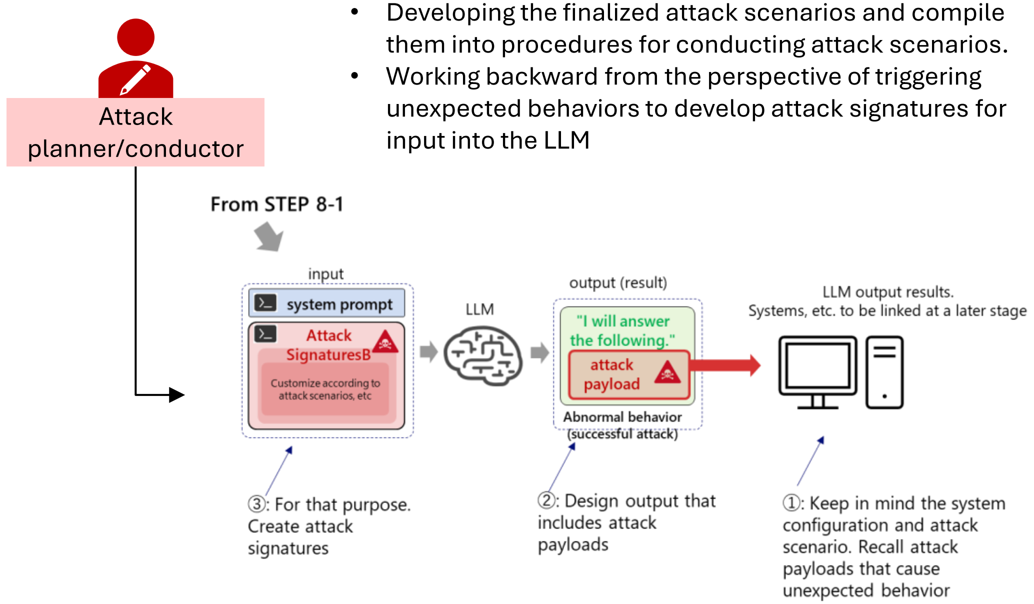
Items

- Based on the results of individual prompt testing, develop the actual attack signatures to be input.
- For each attack scenario, develop a step-by-step execution procedure.
- Procedures for conducting attack scenarios is a structured, reproducible set of steps that outlines the specific attack signatures to be input, environment configurations, and attack execution methods.

Implementation image

Red team

- Developing the finalized attack scenarios and compile them into procedures for conducting attack scenarios.
- Working backward from the perspective of triggering unexpected behaviors to develop attack signatures for input into the LLM


Attack planner/conductor
From STEP 8-1
input
system prompt
Attack SignaturesB
Customize according to attack scenarios, etc
LLM
output (result)
"I will answer the following."
attack payload
Abnormal behavior (successful attack)
LLM output results.
Systems, etc. to be linked at a later stage
③: For that purpose. Create attack signatures
②: Design output that includes attack payloads
①: Keep in mind the system configuration and attack scenario. Recall attack payloads that cause unexpected behavior

Implementation points

- Consider outputs that could introduce risks to the system and develop input attack signatures by reverse-engineering the desired outputs from the perspective of the attack planner/conductor.
- Instead of directly using the attack signatures from individual prompt testing, modifications may be necessary. There is no fixed procedure for making these modifications, making it essential to stay updated on the latest attack techniques and trends.
- An example of developing attack signatures is provided in [Example of deliverables: The report of red teaming results (In Japanese)] *1. In addition, the actual procedures for conducting the attack scenarios is shown in [Example of deliverables: Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)*2.

*1 Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> The report of red teaming results (In Japanese)*2 Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)

3. Explanation of Each Process Process 2

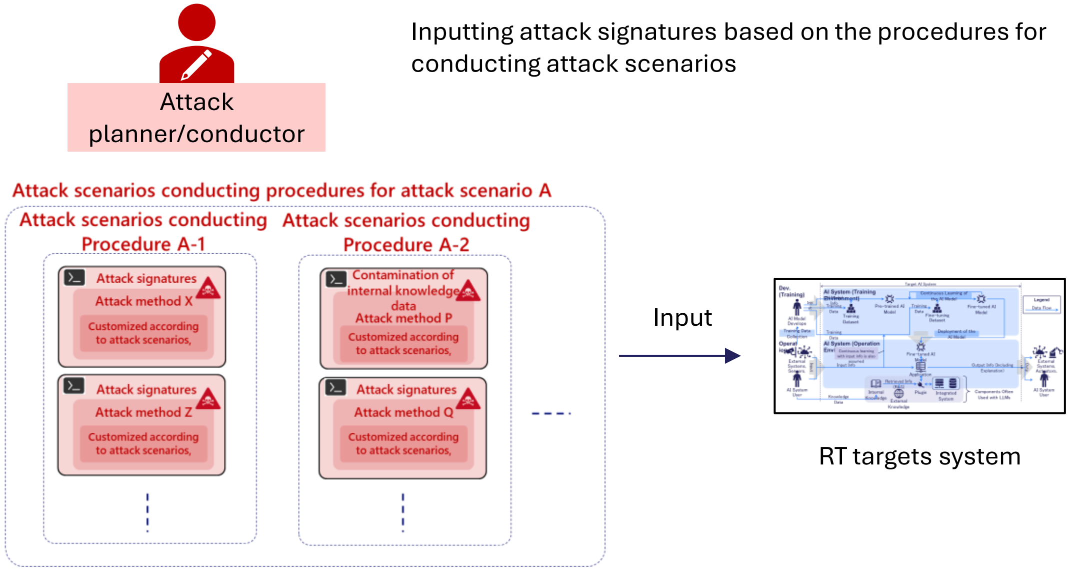
[Details](STEP 8-3) Red teaming for the entire LLM system

Items

- Based on the procedures for conducting attack scenarios developed in **STEP 8-2**, input a series of attack signatures into the target AI system and verify the results.
- Review the output results of the attack signatures, and if necessary, modify them to produce outputs that align with the goals of the attack planner/conductor, then re-input them.
- The attack planner/conductor inputs attack signatures multiple times to refine the attacks.

Implementation image

Red team



Inputting attack signatures based on the procedures for conducting attack scenarios

Attack planner/conductor
Attack scenarios conducting procedures for attack scenario A
Attack scenarios conducting Procedure A-1
Attack signatures
Attack method X
Customized according to attack scenarios,
Attack signatures
Attack method Z
Customized according to attack scenarios,
Attack scenarios conducting Procedure A-2
Contamination of internal knowledge data
Attack method P
Customized according to attack scenarios,
Attack signatures
Attack method Q
Customized according to attack scenarios,
Input
RT targets system

Implementation points

- In LLM systems, behavior is probabilistic and non-deterministic, meaning an attack may succeed after multiple attempts. Therefore, even when using the same attack signature, it is recommended to attempt the attack multiple times.
- The results of the attack scenarios are shown in [Example of deliverables: Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)*.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> Developing risk scenarios, attack scenarios, and results of attack scenarios implementation (In Japanese)

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 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.

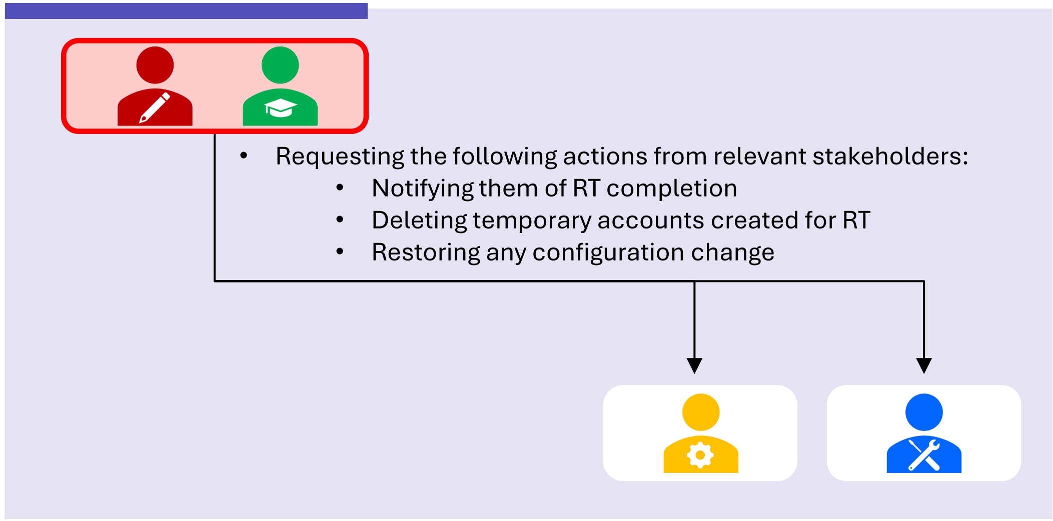
Red team

- 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

3. Explanation of Each Process Process 2

[Details](STEP 9) Record keeping during red teaming

Items

- To preserve detailed evidence of the conducted RT, obtain record during red teaming.
- The collected records should be properly protected following the organization's document management policies and confidential information handling regulations and stored for the designated retention period.

Implementation image

Red team

Attack planner/conductor	AI system expert

Ensuring RT execution records are collected with the following considerations:

- For RT using automation tools, logs should be collected through the tool's logging functionality
- For manual RT, one approach is to set up a proxy along the attack path to capture all passing attack signatures
- If an attack is successful, capture screenshots (screen captures) of the LLM’s output results as evidence

Implementation points

- For RT using automation tools, logs should be collected using the logging functionality of the tool. For manual RT, one approach is to set up a proxy along the attack path to capture all passing attack signatures.
- If an attack is successful, a screenshot (screen capture) should be taken as evidence.

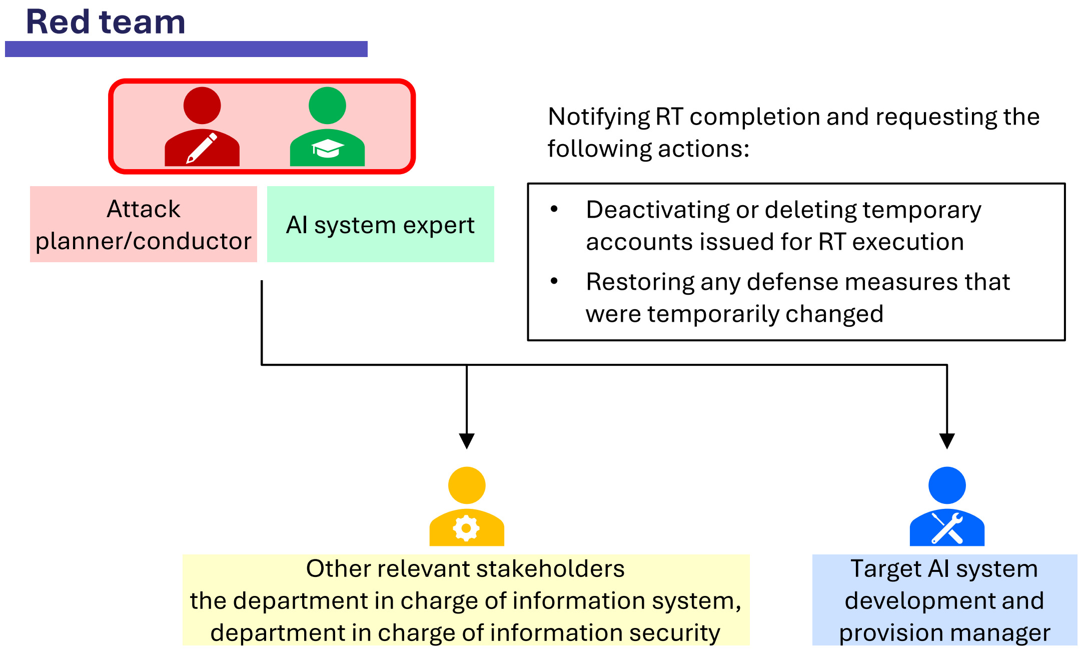
3. Explanation of Each Process Process 2

[Details](STEP 10) After conducting attack scenarios

Items

- Notify relevant stakeholders, such as the target AI system development and provision manager and the department of information systems and information security, after conducting attack scenarios.
- Delete any temporary accounts used for RT. Additionally, if defense measures were temporarily changed, revert them to their original settings.

Implementation image



Red team

Attack planner/conductor

AI system expert

Notifying RT completion and requesting the following actions:

- Deactivating or deleting temporary accounts issued for RT execution
- Restoring any defense measures that were temporarily changed

Other relevant stakeholders
the department in charge of information system,
department in charge of information security

Target AI system development and provision manager

Implementation points

- After RT completion, revert any temporary accounts issued for RT and restore modified settings to their original state.

3. Explanation of Each Process

Process 3: Reporting and Developing Improvement Plans

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

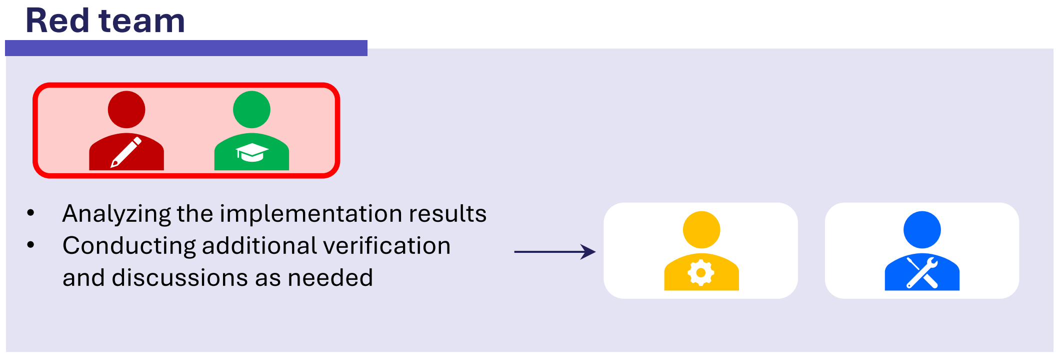
[Legend] ● :Attack planner/conductor ● :AI system expert ● :Target AI system development and provision manager ● :Other relevant stakeholders ● :Business executive officers | Project team | Red team

Process 3: Reporting and Developing Improvement Plans

Red team

STEP 11 Analyzing the red teaming results

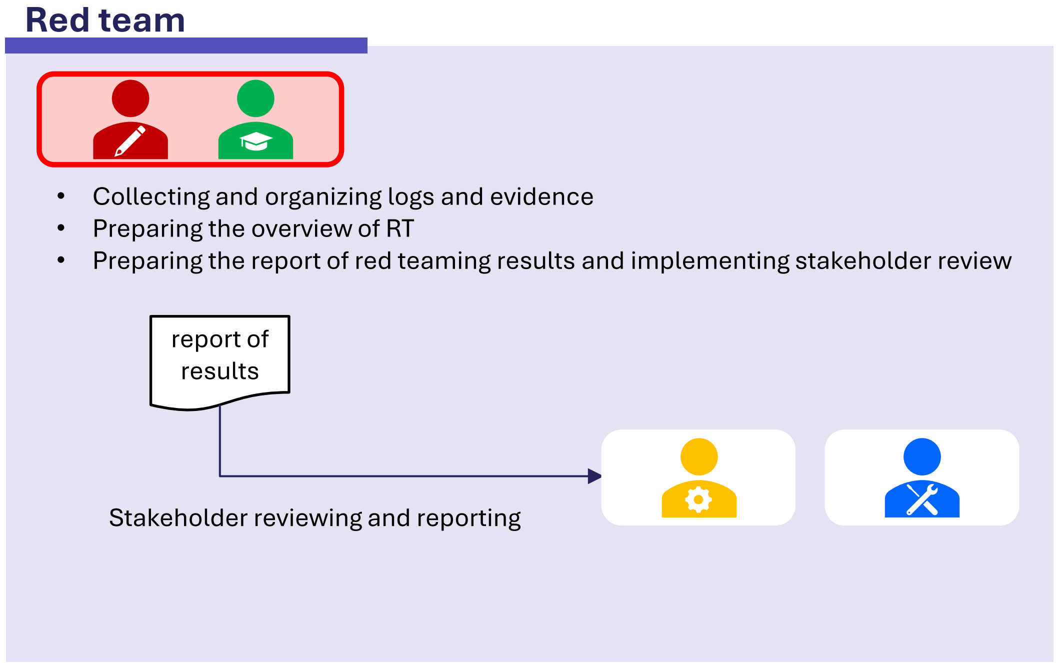
- The attack planner/conductor analyzes the results obtained from red teaming.
- If necessary, additional confirmation of the information to be analyzed is made with relevant departments, such as the development and provision manager of the target AI system, the department in charge of information systems,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



- 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.



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

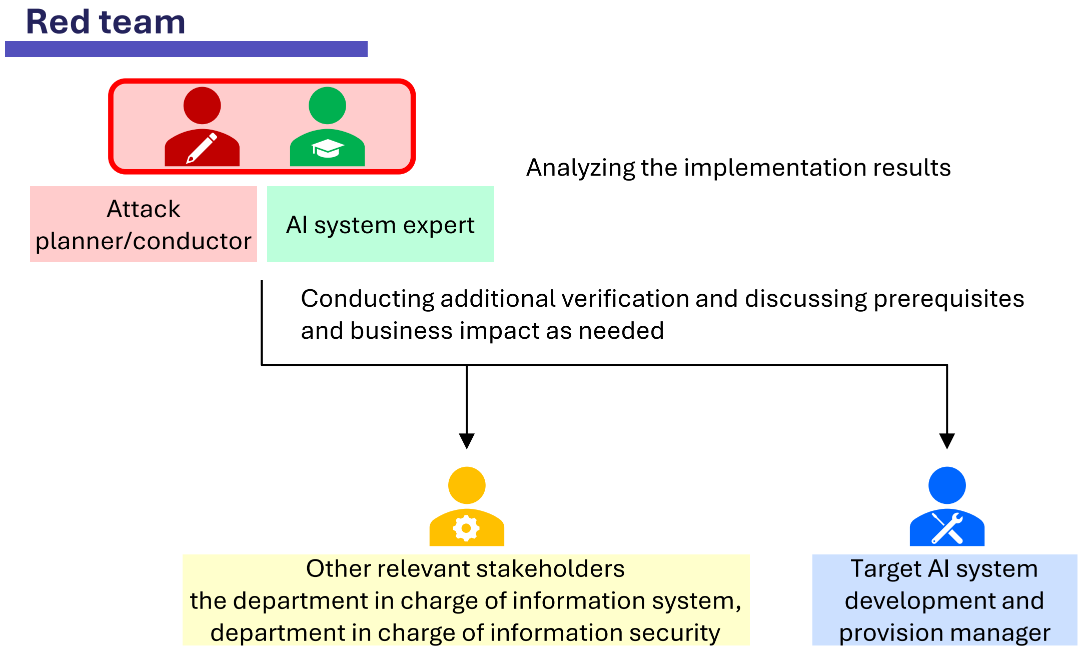
3. Explanation of Each Process Process 3

[Details](STEP 11) Analyzing the red teaming results

Items

- The attack planner/conductor analyzes the results obtained from RT.
- If necessary, the attack planner/conductor may request additional information from relevant departments, such as the target AI system development and provision manager or the department in charge of information system and information security, to support the analysis. Following this, the prerequisites for discovered vulnerabilities are verified, and potential damages, business impact, and necessary countermeasures are discussed.
- If a critical and urgent vulnerability is identified, it must be immediately shared with relevant stakeholders, and countermeasures should be considered.

Implementation image



Red team

Attack planner/conductor

AI system expert

Analyzing the implementation results

Conducting additional verification and discussing prerequisites and business impact as needed

Other relevant stakeholders
the department in charge of information system, department in charge of information security

Target AI system development and provision manager

Implementation points

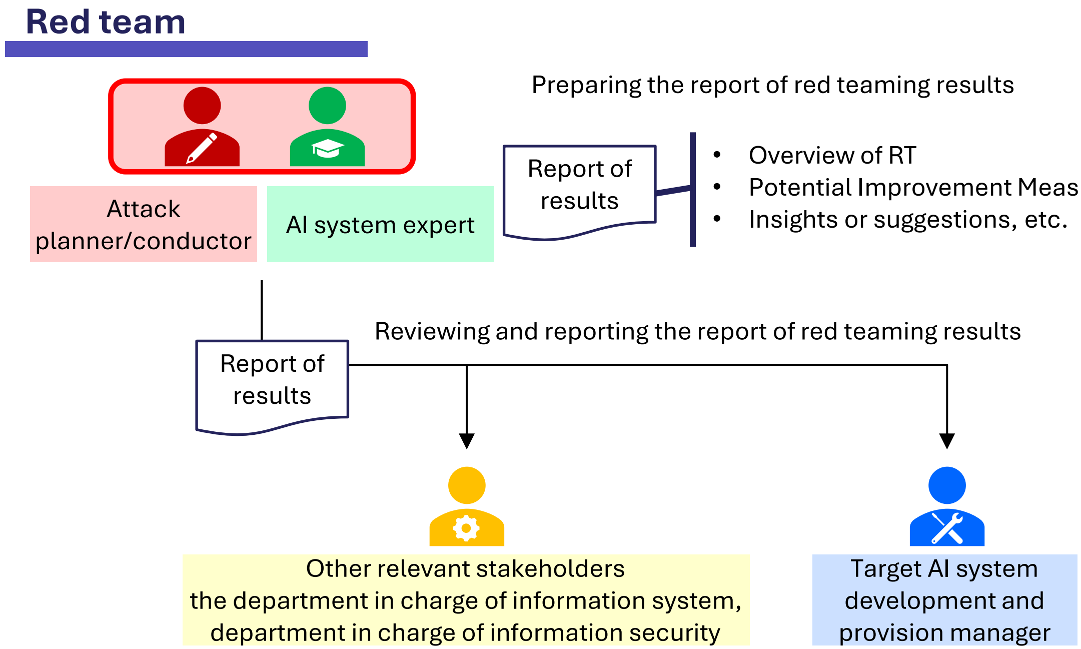
- The attack planner/conductor collaborates with relevant departments and, if necessary, requests additional information to support the analysis. The prerequisites for discovered vulnerabilities are also verified, and an understanding of potential damages and business impact is aligned with stakeholders.

[Details](STEP 12) Preparing the report of red teaming results and implementing stakeholder review

Items

- The attack planner/conductor compiles logs, evidence, and other records based on the discovered vulnerabilities and presents them as an overview of RT.
- The attack planner/conductor then prepares **the report of red teaming results** and conducts a review with the target AI system development and provision manager and other relevant stakeholders to ensure there are no factual inaccuracies.

Implementation image


Red team
Preparing the report of red teaming results
Attack planner/conductor
AI system expert
Report of results
Overview of RT
Potential Improvement Measures
Insights or suggestions, etc.
Reviewing and reporting the report of red teaming results
Report of results
Other relevant stakeholders
the department in charge of information system, department in charge of information security
Target AI system development and provision manager

Implementation points

- At the stage of preparing the report of red teaming results, the focus is on fact-checking the results, rather than analyzing business impact.
- In the report of red teaming results, providing descriptions of the attack signatures used and explaining how the results pose risks or harm helps stakeholders develop a shared understanding more effectively.
- An example of the report of red teaming results is provided in [Example of deliverables: the report of red teaming results (In Japanese)] *.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> the report of red teaming results (In Japanese)

3. Explanation of Each Process Process 3

Process 1 | Process 2 | Process 3
11 | 12 | 13 | 14 | 15

AISI Japan AI Safety Institute

[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

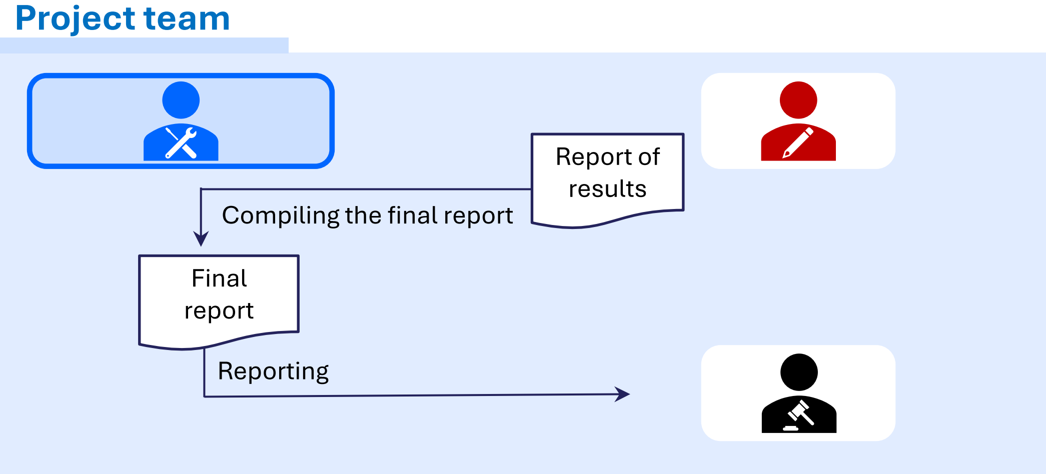
- The development and provision managers of the target AI system should prepare a final report of the red teaming, based on the report of red teaming results reported by the attack planner/conductor.
- If necessary, present the final report to the management team.

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.

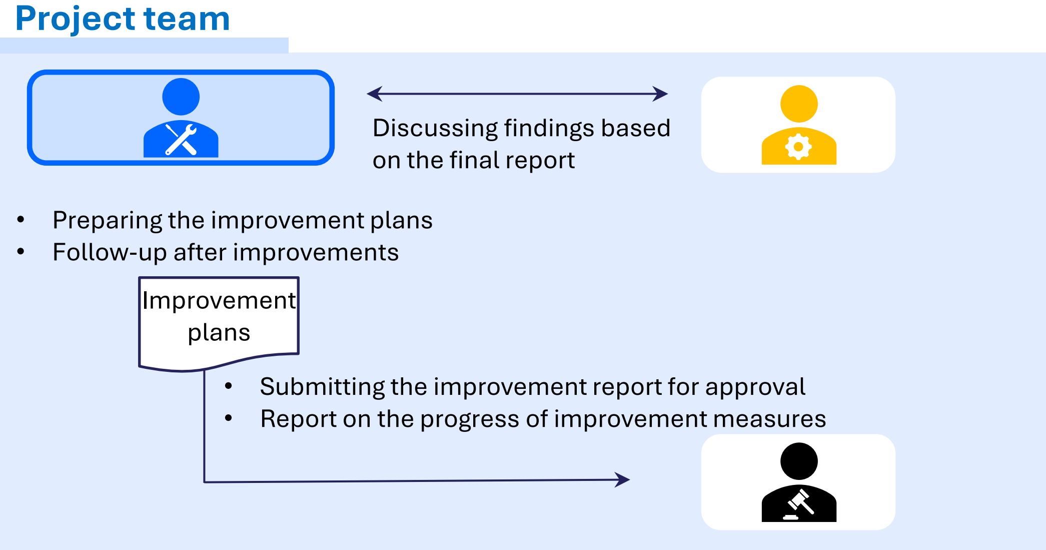

Project team

Report of results

Compiling the final report

Final report

Reporting


Project team

Discussing findings based on the final report

- Preparing the improvement plans
- Follow-up after improvements

Improvement plans

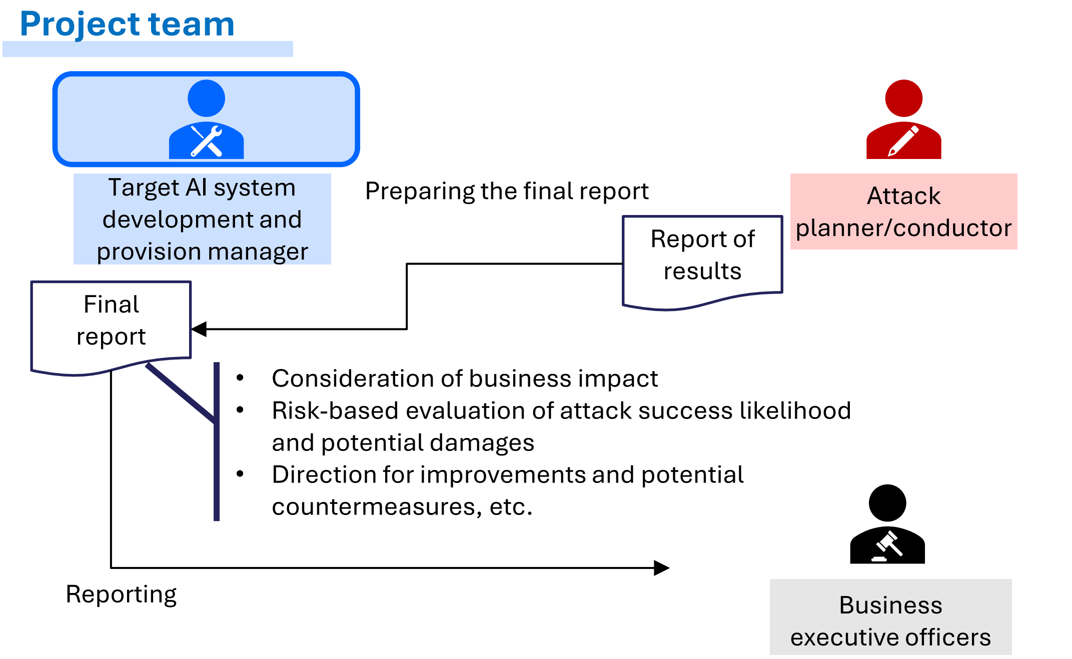
- Submitting the improvement report for approval
- Report on the progress of improvement measures

[Details](STEP 13) Preparing and reporting the final results

Items

- The target AI system development and provision manager prepares the final report based on the report of red teaming results prepared by the attack planner/conductor.
- In the final report, considerations are made regarding business impact based on the system-level risks identified in the report of red teaming results, from the perspective of actual business operations. Subsequently, a risk-based evaluation is conducted on the likelihood of attack success and potential damages.
- If necessary, the final report is presented to business executive officers.

Implementation image


Project team
Target AI system development and provision manager
Preparing the final report
Attack planner/conductor
Report of results
Final report
Reporting
Business executive officers

- Consideration of business impact
- Risk-based evaluation of attack success likelihood and potential damages
- Direction for improvements and potential countermeasures, etc.

Implementation points

- While the purpose of the report of red teaming results is fact verification, the purpose of the final report is to analyze business impact from an actual business perspective and conduct a risk-based evaluation.
- The metrics for risk-based evaluation can be referenced from the overall risk evaluation metrics considered during developing risk scenarios.
- An example of the final report is provided in [Example of deliverables: the final report (In Japanese)] *.

* Guide to Red Teaming Methodology on AI Safety>Supplementary Document(Example of deliverables)> the final report (In Japanese)

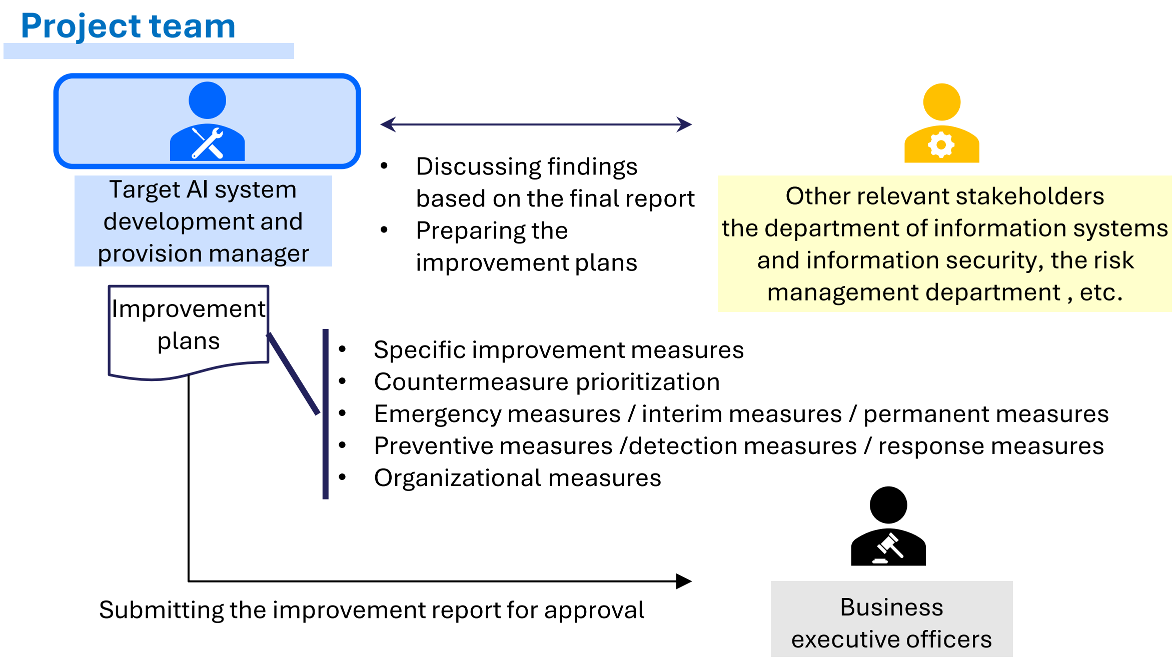
3. Explanation of Each Process Process 3

[Details](STEP 14) Developing and implementing improvement plans

Items

- The target AI system development and provision manager considers business risks and other factors to develop specific improvement measures and prepare the improvement plans.
- When considering specific improvement measures and the improvement plans, priorities should be determined based on urgency and risk level.
- After obtaining approval from business executive officers, the improvement plans are finalized.

Implementation image



Project team

Target AI system development and provision manager

- Discussing findings based on the final report
- Preparing the improvement plans

Other relevant stakeholders the department of information systems and information security, the risk management department , etc.

Improvement plans

- Specific improvement measures
- Countermeasure prioritization
- Emergency measures / interim measures / permanent measures
- Preventive measures /detection measures / response measures
- Organizational measures

Submitting the improvement report for approval

Business executive officers

Implementation points

- When preparing the improvement plans, prioritize the improvement measures listed in the final report based on urgency and risk level, considering their business impact.
- Consider not only system-related improvements but also organizational measures, such as revising operational processes.

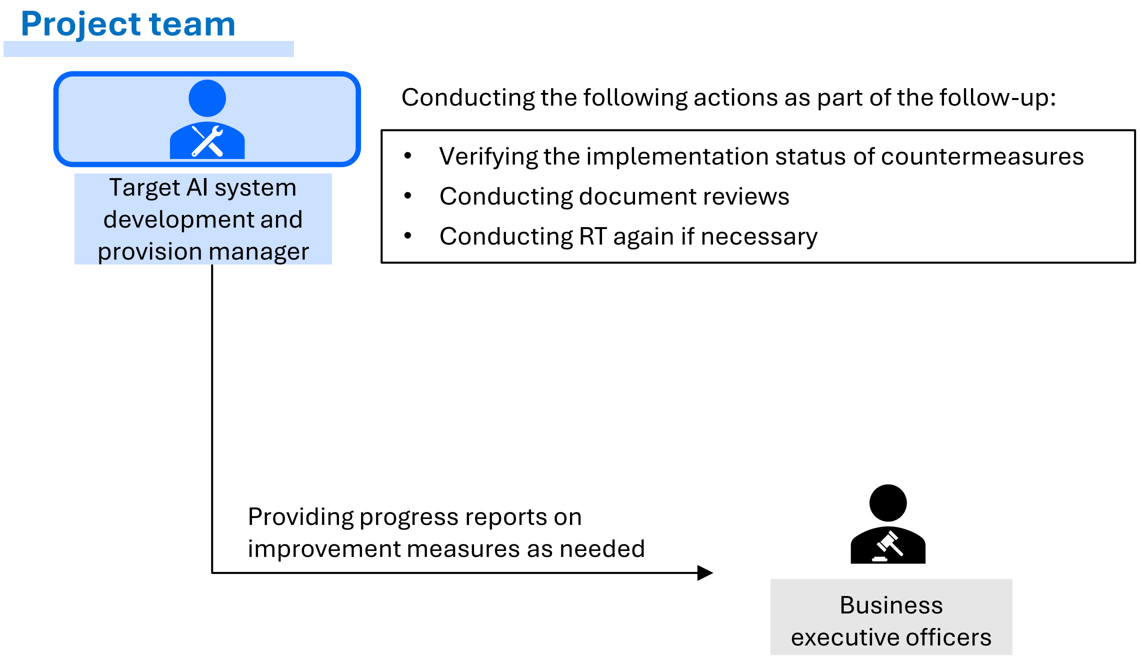
3. Explanation of Each Process Process 3

[Details](STEP 15) Follow-up after improvement

Items

- The progress of improvement measures implemented based on the improvement plans should be periodically reviewed in executive meetings or other relevant forums.
- After implementing the improvement measures, it is recommended to verify the status of countermeasure settings, conduct document reviews, or, if necessary, conduct RT again. Subsequently, it is advisable to confirm that the identified vulnerabilities have been properly addressed and that the associated risks have been mitigated.

Implementation image



Project team
Target AI system development and provision manager
Business executive officers

Conducting the following actions as part of the follow-up:

- Verifying the implementation status of countermeasures
- Conducting document reviews
- Conducting RT again if necessary

Providing progress reports on improvement measures as needed

Implementation points

- By ensuring strict progress management and conducting RT again, if necessary, a continuous improvement cycle can be maintained, enabling the promotion of effective and practical improvement activities.

AISI

Japan AI Safety Institute