

Guide to Evaluation Perspectives on AI Safety (Version 1.10)

Summary

**AI Safety Institute
(March 28, 2025)**



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 AI Safety Evaluations

... P.6

5. Key Elements of AI Safety

... P.7

6. Evaluation Perspectives on AI Safety

... P.8

7. Evaluator and Evaluation Timing

... P.11

**Revision of the Guide to Evaluation
Perspectives on AI Safety**

... P.13

1. Background and Purpose

With the rapid progress of AI systems, AI Safety is becoming increasingly important. This document has been prepared to present basic concepts on AI Safety evaluations.

Background

- The development and widespread adoption of AI-related technologies have been rapid throughout society. In addition, the emergence of generative AI, especially the foundation model accelerates innovation. At the same time, **concerns are growing both domestically and internationally about so-called AI Safety**, including the malicious use or misuse of AI systems and concerns about inaccurate outputs.
- Japan has led the Hiroshima AI Process towards the realization of safe, secure and reliable AI, and has contributed to the formulation of global rules related to AI safety, such as compiling the Hiroshima Process International Guiding Principles.
- **AI safety is a prerequisite for the rapid progress of AI to contribute to the sustainable development of society.**

Purpose

- This Guide to Evaluation Perspectives on AI Safety (hereinafter referred to as “this document”) presents basic concepts that those involved in the development and provision of AI systems can refer to when conducting AI Safety evaluations. Specifically, this document provides the following:
 - ✓ Perspectives of AI Safety evaluations, examples of possible risks, examples of evaluation items
 - ✓ Ideas on who will conduct the evaluation and when it will be conducted
 - ✓ Summary of evaluation method

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 consideration of the AI Guidelines for Business published in April 2024, this document has been prepared based on a survey of international publications and tools.

International Publications

Artificial Intelligence Risk Management Framework (AI RMF 1.0)

A document that specifies a framework to help facilitate the responsible design, development, deployment, and use of AI systems

AI 600-1: Generative Artificial Intelligence Profile

A document to help identify inherent risks posed by generative AI and suggest actions for optimal generative AI risk management.

CATALOGUING LLM EVALUATIONS

A document describing the taxonomy, future issues, and methodology (recommended assessment and testing approaches) for LLM assessments.

Model AI Governance Framework for Generative AI

A framework for international consensus on the governance of generative AI, based on the Model AI Governance Framework.

International Scientific Report on the Safety of Advanced AI: interim report

The document organizes the latest information on advanced AI capabilities and risks for discussion at the AI Seoul Summit, co-hosted in May 2024.

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 Evaluation Perspectives on AI Safety

Tools (Organization)

Robust Intelligence Platform (Robust Intelligence)

Tools that can automatically ensure security through real-time protection and testing during development and operation of AI models.

Citadel (Citadel AI)

The tool speeds up automatic validation and quality improvement by testing and monitoring AI models during training and operation.

Project Moonshot (AI Verify Foundation)

An open source LLM evaluation tool developed by the AI Verify Foundation in Singapore.

Inspect (AI Safety Institute, UK)

An open-source evaluation tool dedicated to LLM for AI systems.

LLM Observability (Arize)

The tool can be used for-automatic monitoring and evaluating LLMs in operation and focuses on visualizing the status of AI systems.

3. Structure of This Document

The basic concepts that can be referred to when conducting an AI Safety evaluations are categorized by type. The table of contents is organized for easy reference and classification are listed.

- The contents of each section of this document are described based on the items organized from a 5W1H perspective.
- The main intended audience is AI developers and AI providers. In particular, those managers and executives.

Type	Examples of items to be described
What (What is evaluation? What to evaluate?)	<ul style="list-style-type: none"> ➤ AI systems covered in this document ➤ Definition and scope of "evaluation" on AI safety ➤ Evaluation perspectives on AI Safety
Why (Why do we value it?)	<ul style="list-style-type: none"> ➤ Purpose and Significance of AI Safety evaluations
Who (Who evaluates?)	<ul style="list-style-type: none"> ➤ What role will the person(s) play in conducting the evaluation?
When (When to evaluate?)	<ul style="list-style-type: none"> ➤ Evaluation timing
Where (Where to evaluate?)	<ul style="list-style-type: none"> ➤ Whether it is conducted by own organization or by a third party (an external organization conducting the evaluation)
How (How to evaluate?)	<ul style="list-style-type: none"> ➤ Evaluation method (technical evaluation and managerial evaluation)

Guide to Evaluation Perspectives on AI Safety [Table of Contents]	
1	Introduction
2	AI Safety
3	Details of Evaluation Perspectives
4	Evaluator and the Evaluation Timing
5	Evaluation Methods
6	Considerations for Evaluation
A	Appendix

Intended Audience

AI Developers and AI Providers Development and Provision Managers Business Executives Officers

4. Scope of AI Safety Evaluations

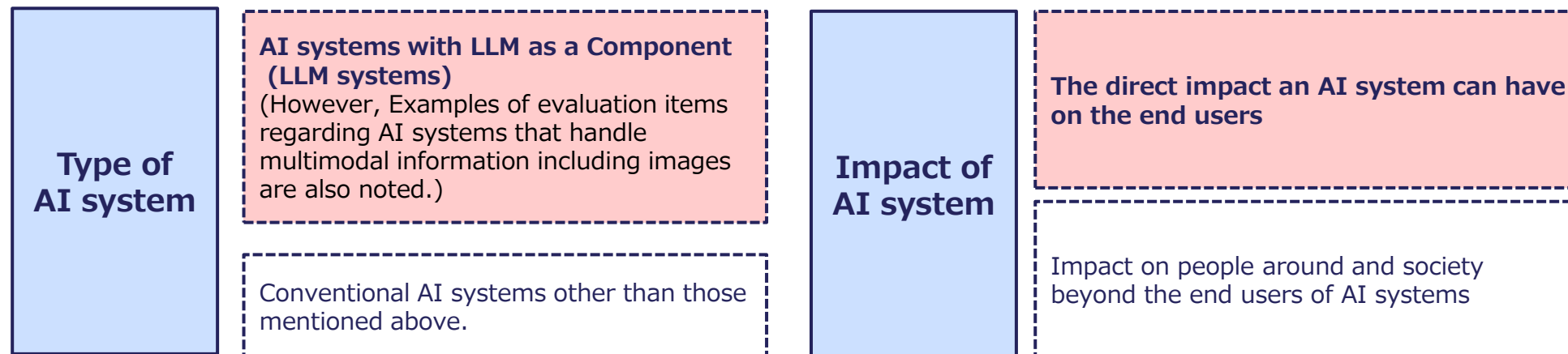
AI safety evaluations in this document describe as the determination of whether an AI system is appropriate in terms of AI Safety perspective. The evaluation perspectives described here are specifically focused on AI systems(LLM systems) that incorporate large language models.

🔍 Scope of AI Safety evaluations in this document

- **AI safety evaluations is “determination of whether an AI system is appropriate in terms of AI Safety perspective.”**
The AI safety perspective consists of "human-centric," "safety," "fairness," "privacy protection," "ensuring security," and "transparency" as key elements.
- The scope of the AI safety evaluations in this document is organized in terms of **(1) Type of AI system** and **(2) Impact of AI system**.

Scope of AI Safety Evaluation in this document

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









*Regarding the evaluation perspectives for AI systems including foundation models that handle multimodal information, additional entries have been included in the examples of evaluation items, with a focus on cases involving image processing. Future revisions will be considered based on technological and usage trends. The points to note regarding this are detailed in "Considerations for Evaluation" section.

5. Key Elements of AI Safety

For improving AI safety, key elements to emphasize include “Human-Centric,” “Safety,” “Fairness,” “Privacy Protection,” “Ensuring Security,” and “Transparency.”

- Among the items outlined in the "C. Common Guiding Principles" section of the AI Guidelines for Business, the following six items are identified as key elements that should be prioritized to enhance AI Safety*.
- This document derives AI Safety evaluation perspectives related to these key elements.

Key Elements	Brief Explanation
(1) Human-Centric 	During the development, provision, and use of an AI system and service, the human rights guaranteed by the Constitution or granted internationally should not be violated, as the foundation for accomplishing all matters to be conducted. In addition, an action should be taken in a way that AI expands human abilities and enables diverse people to seek diverse well-being.
(2) Safety 	During the development, provision, and use of an AI system and service, damage to the lives, bodies, or properties of stakeholders should be avoided. In addition, damage to the minds and the environment should be avoided.
(3) Fairness 	During the development, provision, and use of an AI system and service, efforts should be made to eliminate unfair and harmful bias and discrimination against any specific individuals or groups based on race, gender, national origin, age, political opinion, religion, and other diverse backgrounds. In addition, before developing, providing, or using AI systems or services, each entity should recognize that there are some unavoidable biases even if such attention is paid, and determines whether the unavoidable biases are allowable from the viewpoints of respect for human rights and diverse cultures.
(4) Privacy Protection 	During the development, provision, and use of an AI system and service, privacy should be respected and protected in accordance with its importance. At the same time, relevant laws should be obeyed.
(5) Ensuring Security 	During the development, provision, and use of an AI system and service, security should be ensured to prevent the behaviors of AI from being unintentionally altered or stopped by unauthorized manipulations.
(6) Transparency 	During the development, provision, and use of an AI system and service, based on the social context when the AI system or service is used, information should be provided to stakeholders to the reasonable extent necessary and technically possible while ensuring the verifiability of the AI system or service.

* Of the issues to be addressed by entity, "Accountability" is ensured by confirming measures for the other six, distributing and clearly stating the legal and practical responsibilities of each stakeholder to a reasonable extent, and collecting and disclosing appropriate information as necessary.

6. Evaluation Perspectives on AI Safety

Based on the results of various surveys and other sources, the perspectives for AI safety evaluations have been organized.

- Considering the descriptions in the AI Guidelines for Business, international publications, and survey on tools, the evaluation perspectives related to the key elements of AI Safety have been organized.

		Evaluation perspectives on AI safety									
		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
Key Elements of AI Safety	Human-centric	●	●	●	●						
	Safety	●	●		●				●	●	
	Fairness	●		●						●	
	Privacy protection					●					
	Ensuring security						●				
	Transparency		●	●					●	●	●

※ Various studies on AI Safety evaluations are ongoing domestically and internationally across diverse fields in industry, government, and academia, and the situation is constantly changing. Therefore, this document presents the evaluation perspectives that are considered to be particularly important. The perspectives described in this document are not exhaustive and are expected to be updated in the future.

A summary of expected goals for each evaluation perspective

This document presents evaluation perspectives on AI safety considering recent technological trends.

	Evaluation perspectives on AI safety	Relevant key elements	Expected goals (After effective measures are implemented)
(1)	Control of Toxic Output	Human-Centric, Safety, Fairness	<ul style="list-style-type: none"> LLM system can control the output of harmful information, such as information about terrorism and crime or offensive expressions.
(2)	Prevention of Misinformation, Disinformation and Manipulation	Human-Centric, Safety, Transparency	<ul style="list-style-type: none"> Fact-finding mechanism for LLM system outputs is placed. Manipulation of end user's decisions by the output of the LLM system is avoided.
(3)	Fairness and Inclusion	Human-Centric, Fairness, Transparency	<ul style="list-style-type: none"> The output of the LLM system does not contain harmful biases and is free from unfair discrimination against any individual or group. The output of the LLM system is understandable, i.e., highly readable, to all end users.
(4)	Addressing High-risk Use and Unintended Use	Human-Centric, Safety	<ul style="list-style-type: none"> No harm or disadvantage is caused by inappropriate use of the LLM system that deviates from its intended purpose.
(5)	Privacy Protection	Privacy Protection	<ul style="list-style-type: none"> LLM system protects privacy appropriately according to the importance of the data it handles.

A summary of expected goals for each evaluation perspective

This document presents evaluation perspectives on AI safety considering recent technological trends.

	Evaluation perspectives on AI safety	Relevant key elements	Expected goals (After effective measures are implemented)
(6)	Ensuring Security	Ensuring Security	<ul style="list-style-type: none"> Leakage of confidential information due to unauthorized operations and unintended modification or shutdown of the LLM system are prevented.
(7)	Explainability	Transparency	<ul style="list-style-type: none"> The basis for the output can be confirmed to a technically reasonable extent for the purpose of presenting evidence of the LLM system's operation, etc.
(8)	Robustness	Safety, Transparency	<ul style="list-style-type: none"> LLM system provides stable output against unexpected inputs such as adversarial prompting, garbled data, and erroneous input.
(9)	Data Quality	Safety, Fairness, Transparency	<ul style="list-style-type: none"> The data accessed by LLM systems are in an appropriate state, including during model training, and that the history of the data is properly managed.
(10)	Verifiability	Transparency	<ul style="list-style-type: none"> Various types of verification against LLM system are available from the model learning phase and the development/provision phase of the LLM system to the time of use.

7. Evaluator and Evaluation Timing

The AI safety evaluations are basically conducted by the development and provision managers in AI development and AI Provision. In addition, the AI safety evaluations shall be repeated within a reasonable range and at an appropriate timing.

Evaluator

- The primary conductor of AI Safety evaluation are **development and provision managers in AI development and AI provision**.
- The role will be performed by which person depends on **the life cycle of the AI system**.
- In order to provide independence in objective evaluation and decision-making regarding system development and provision, **evaluation by experts in the own/other organization** who are not directly involved in the development and provision of the target system and third party can also be effective.

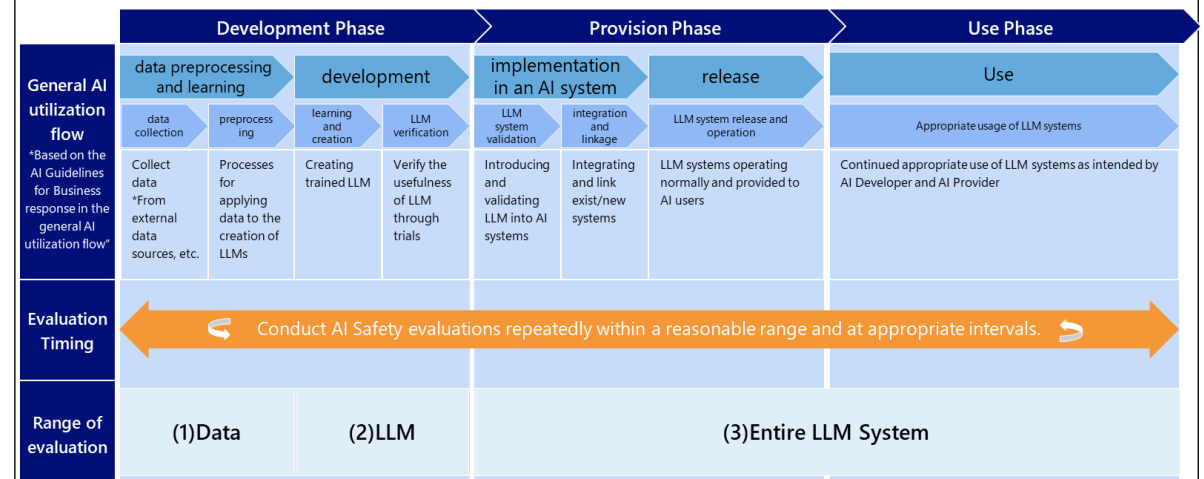
Evaluator of AI Safety

Evaluators according to lifecycle	Development manager in AI development: Data learning and model building stages related to AI systems.
	Provision manager in AI provision: Stage of integrating AI systems into applications, etc.
Type of Evaluators	Own organization Those directly involved in the development and provision of AI systems
	Own/other organization Experts from own/other organization (not directly involved in the development or provision of the AI system)
	Other organization Third parties (external organizations conducting independent evaluation)

Evaluation Timing

- The timing of AI safety evaluations should be within a reasonable range and conducted at appropriate intervals during the phases of development, provision, and use of the AI system.
- AI safety evaluations will be **repeated, not just once**.
- Depending on the phases of development, provision, and use, the **range of evaluation will differ**.

Evaluation Timing in the flow to utilize LLM system



A high-level mapping between the evaluation perspectives outlined in this document and those described in international Publications.

● : There are descriptions related to this perspective in international Publications.

Evaluation perspectives on AI safety	Artificial Intelligence Risk Management Framework (AI RMF 1.0)	AI 600-1: Generative Artificial Intelligence Profile	Cataloguing LLM Evaluations	Model AI Governance Framework for Generative AI	International Scientific Report on the Safety of Advanced AI (Interim report)
① Control of Toxic Output		●	●		●
② Prevention of Misinformation, Disinformation and Manipulation	●	●		●	●
③ Fairness and Inclusion	●	●	●	●	●
④ Addressing High-risk Use and Unintended Use		●			●
⑤ Privacy Protection	●	●		●	●
⑥ Ensuring Security	●	●	●	●	
⑦ Explainability	●		●	●	●
⑧ Robustness	●		●		●
⑨ Data Quality	●	●	●	●	●
⑩ Verifiability		●			●

Please note that the mapping reflects the state at the time of this document's publication and is subject to change.

Revision of the Guide to Evaluation Perspectives on AI Safety

AI Safety Institute
(March 28, 2025)

This document has been revised to accommodate AI safety evaluations across a wider range of target domains, and to serve as a more concrete guide for specific evaluation items.

Background

- In September 2024, this document was created with the aim of presenting fundamental principles that can be referenced by those involved in the development and provision of AI systems when conducting AI safety evaluations.
- While this document primarily focuses on LLM systems^{*1}, considering recent global trends—such as Big Tech beginning to support multimodal^{*2} foundation models^{*3*4} that include components like images, there is an increasing demand for AI safety evaluations that extend beyond AI systems solely incorporating LLMs to encompass a broader range of diverse AI systems.

*1: AI systems that incorporate generative AI as a component, particularly those that include LLMs as part of their structure.

Reference: https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/

*2: This document specifies the types of information and the formats of input data.

*3: (Foundation Model): An AI model trained on extensive data and capable of being adapted to a wide range of downstream tasks.

*4: (Multimodal Foundation Model): A foundation model that associates and processes information based on multiple modalities that mediate human communication and sensory experiences.

Purpose

- Investigate AI safety evaluation perspectives and identify examples of evaluation criteria for each perspective in the context of assessing multimodal foundation models, ultimately guiding potential revisions to this document

Investigated on academic papers addressing threats and risks related to AI safety, as well as AI safety evaluation methods, in AI systems that incorporate multimodal foundation models as components.

【Investigation method】

Preliminary investigation	Target AI systems	Multimodal foundation models
	Target period	Academic papers registered in information sources after 2022
	Information sources	Pre-print server (e.g., arXiv) 、 Major international conferences (e.g., AAAI, ECCV, CVPR, EMNLP, ICCV, ICML, ICRA, IJCAI, ICLR, KDD, NeurIPS) and Academic journals (e.g., IEEE Transactions on Information Forensics and Security, IEEE Transactions on Pattern Analysis and Machine Intelligence, Journal of Medical Internet Research)
	Investigation method	Search the above information source by using relevant keywords (AI Safety, Multimodal, Security) , and extracted the documents. As the preliminary review of the extracted documents, the abstracts were examined, along with verification of whether these documents include evaluation perspectives related to AI safety in multimodal foundation models or not.



Mainly extracted documents which describes risks and investigation methods regarding relevant each investigation perspective for AI Safety.

Follow-up Investigation

Based on the content of the document targeted for follow-up investigation, extracted key evaluation perspectives and examples of evaluation criteria which become important for AI safety evaluation of multimodal foundation models and added them to the current guide.

Based on the results of the investigation, additional perspectives requiring particularly careful evaluation and example evaluation items were appended to the existing guidelines.

- Through this investigation, examples of evaluation items important for assessing AI systems that primarily handle multimodal information, such as images, were organized as follows and appended to the corresponding evaluation perspectives in Chapter 3. *This includes examples within the scope of current guidelines (e.g., LLM systems).

Evaluation perspectives	Examples of Evaluation Items
Control of Toxic Output	Text or images may appear harmless when viewed in isolation, but their combination can sometimes have a disturbing effect. Can harmful output still be prevented in response to such input? When an image or text input contains false, nonsensical, or highly complex content, the model's behavior may sometimes become unstable. Can harmful output still be prevented in such cases?
Fairness and Inclusion	When asked to generate images of individuals related to a specific category (e.g., occupation), does the output remain free from biases related to personal attributes such as gender, age, or race?
Privacy Protection	Can the model refuse to respond when asked for personally identifiable information (PII) along with visual clues about individuals (e.g., a photo of a person's face)?
Ensuring Security	Even if defensive measures are effective against prompt injection attacks using text input, they may not be as effective when the same text is embedded in an image. Is it possible to defend appropriately against this type of input? Can responses be refused when inputs prompt the execution or generation of code that could potentially result in harmful outcomes? Can the system determine whether combined text and image inputs, individually or collectively, are adversarial? For example, by using thresholds to classify harmless and adversarial images?
Robustness	Does the system operate reliably even when perturbed data is input, such as images with noise causing blurriness, adversarial images, images containing harmful text, or data with perturbations in both images and text? In some cases, models may behave unstably when an image unrelated to the question or prompt is input. Do the models remain stable even in such cases? If an image is paired with text containing false information, can the system still generate a response that aligns with the facts? When a complex and perplexing image (e.g., an optical illusion that causes visual confusion) is input, the model's behavior may become unstable. Does it remain stable in such cases? For information that is highly related and tends to appear together but has no actual direct causal relationship (e.g., pacifiers often appear together with infants), recognition accuracy may drop when the related information is absent (e.g., the accuracy of pacifier recognition decreases in images without infants). Can stable recognition still be achieved in such cases without being influenced by peripheral information? Even if the image contains noise such as motion blur (blurring caused by moving subjects) or occlusion (phenomena where objects are hidden by other objects), does the output maintain accuracy in response to text instructions about the content of an input image?
Data Quality	Are the quantity and quality of the training data sufficient to ensure the system is trained to a practical standard.

AISI

Japan AI Safety Institute