Guide to Evaluation Perspectives on AI Safety (Version 1.00)

Summary

Japan AI Safety Institute (September 18, 2024)

AISI Japan
AI Safety Institute

Table of Contents

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 evaluation.

AISI Japan AI Safety Institute

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 evaluations on AI Safety. Specifically, this document provides the following:
  - ✓ Evaluation perspectives on AI Safety, 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:
"State that maintained safety and fairness to reduce societal risks* arising from AI use, privacy protection to prevent of inappropriate use of personal data, ensuring security against risks such as external attack caused by vulnerabilities of AI systems, and transparency by ensuring the verifiability of systems and providing appropriate information, based on the human-centric concept."

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

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.

AISI Japan AI Safety Institute

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.

International Publications

Tools (Organization)

Guide to Evaluation Perspectives on AI Safety

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.

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 accelerates 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 evaluations on AI Safety are categorized by type. The table of contents is organized for easy reference and classification are listed.

AISI Japan AI Safety Institute

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

Type	Examples of Items Described in This Document
What (What is evaluation? What to evaluate?)	➢ 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?)	➢ Purpose and significance of evaluations on AI Safety
Who (Who evaluates?)	➢ What role will the person(s) play in conducting the evaluation?
When (When to evaluate?)	➢ Evaluation timing
Where (Where to evaluate?)	➢ Whether it is conducted by own organization or by a third party (an external organization conducting the evaluation)
How (How to evaluate?)	➢ Evaluation method (technical evaluation and managerial evaluation)

Intended Audience

AI Developers and AI Providers | Development and Provision Managers | Business Executives Officers
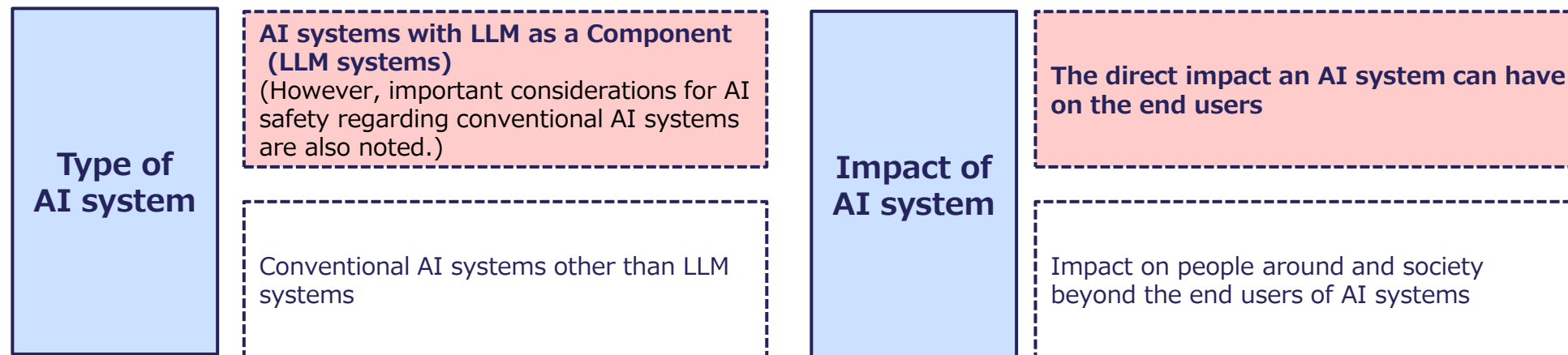
4. Scope of Evaluations on AI Safety

Evaluations on AI Safety 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 Evaluations on AI Safety in This Document

➤ **AI Safety evaluation** 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 evaluations on AI Safety in this document is organized in terms of **(1) Type of AI system** and **(2) Impact of AI system.**

Scope of Evaluations on AI Safety in this document
The boxes noted in red below indicate the scope of the evaluation in this document.

Type of AI system	**AI systems with LLM as a Component (LLM systems)** (However, important considerations for AI safety regarding conventional AI systems are also noted.)
	Conventional AI systems other than LLM systems

Impact of AI system	**The direct impact an AI system can have on the end users**
	Impact on people around and society beyond the end users of AI systems

*The evaluation perspectives for AI systems including foundation models that handle multimodal information will be considered in the future, taking into account technological and usage trends. The points to note regarding this are detailed in "Considerations for Evaluation" section.

6

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

AISI Japan AI Safety Institute

- 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 evaluation perspectives on AI Safety 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.

Based on the results of various surveys and other sources, the evaluation perspectives on AI Safety have been organized.

AISI | Japan AI Safety Institute

- 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 to 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	• 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	• 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	• 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 to High-risk Use and Unintended Use**	Human-Centric, Safety	• No harm or disadvantage is caused by inappropriate use of the LLM system that deviates from its intended purpose.
(5)	**Privacy Protection**	Privacy Protection	• LLM system protects privacy appropriately according to the importance of the data it handles.

A summary of expected goals for each evaluation perspective

AISI Japan AI Safety Institute

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	• Leakage of confidential information due to unauthorized operations and unintended modification or shutdown of the LLM system are prevented.
(7)	**Explainability**	Transparency	• 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	• LLM system provides stable output against unexpected inputs such as adversarial prompting, garbled data, and erroneous input.
(9)	**Data Quality**	Safety, Fairness, Transparency	• 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	• 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 evaluations on AI Safety are basically conducted by the development and provision managers in AI development and AI Provision. In addition, the evaluations on AI Safety shall be repeated within a reasonable range and at an appropriate timing.

AISI | Japan AI Safety Institute

👤 Evaluator

- The primary conductor of evaluations on AI Safety 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.
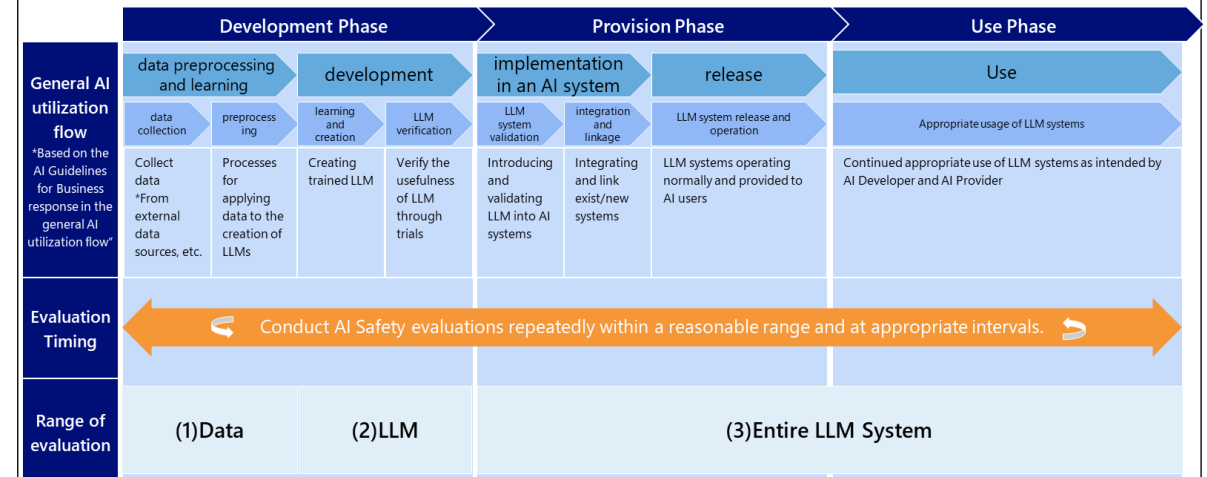
Evaluator of AI Safety

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 evaluations on AI safety should be within a reasonable range and conducted at appropriate intervals during the phases of development, provision, and use of the AI system.
- Evaluations on AI safety 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

General AI utilization flow *Based on the AI Guidelines for Business response in the general AI utilization flow"	Development Phase				Provision Phase			Use Phase
	data preprocessing and learning		development		implementation in an AI system		release	Use
	data collection	preprocessing	learning and creation	LLM verification	LLM system validation	integration and linkage	LLM system release and operation	Appropriate usage of LLM systems
	Collect data *From external data sources, etc.	Processes for applying data to the creation of LLMs	Creating trained LLM	Verify the usefulness of LLM through trials	Introducing and validating LLM into AI systems	Integrating and link exist/new systems	LLM systems operating normally and provided to AI users	Continued appropriate use of LLM systems as intended by AI Developer and AI Provider
Evaluation Timing	Conduct AI Safety evaluations repeatedly within a reasonable range and at appropriate intervals.							
Range of evaluation	(1)Data		(2)LLM		(3)Entire LLM System			

References
High-level mapping of the evaluation perspectives described in this document with those described in international publication

AISI Japan AI Safety Institute

● : International publication that has description related to those perspectives.

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 to 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 high-level mapping is at the time of publication of this document, and that may subject to change.

AISI

Japan AI Safety Institute