Introduction to “Guide to Evaluation Perspectives on AI Safety” and “Guide to Red Teaming Methodology on AI Safety

AI Safety Institute

AISI Japan AI Safety Institute

AISI Japan AI Safety Institute

AISI Japan AI Safety Institute

Increase in Businesses Providing Systems/Services with AI Capabilities

- Generative AI has become widespread, being used for text summarization, researches, data analysis. In addition, various AI functionalities are now accessible via open-source platforms and commercial APIs, enabling business to integrate AI into their systems or services.
- This document introduces the key considerations for “AI Providers” when developing and offering AI systems or services and provides an overview of guidelines developed in AISI.

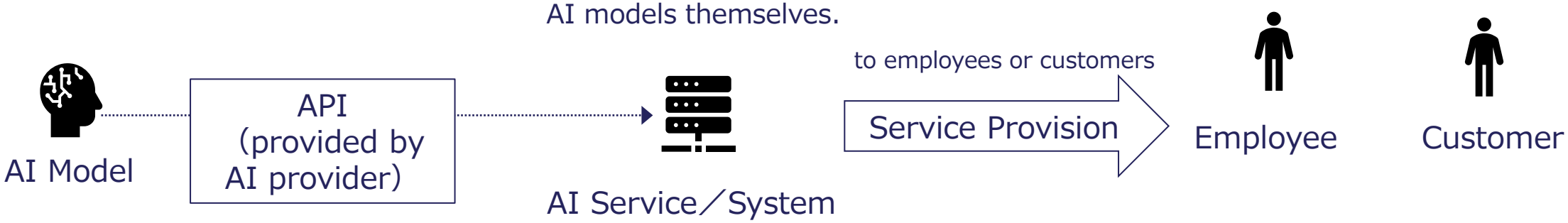
	Summary	Example
AI Business Users	Businesses using AI systems/services.	✓ Summarising research papers with generative AI. ✓ Generating Python code with AI.
AI Providers	Businesses using AI systems/services.	✓ Using LLM APIs or open-source LLMs to develop AI systems ✓ Creating chatbots with generative AI APIs for internal or external use.
AI Developers	Businesses using AI systems/services.	✓ Building AI models, algorithms, training, and system architecture.

Note: Please reference AI Business Guidelines（https://aisi.go.jp/effort/effort_information/250328_2/）for the definitions for AI Users, Providers, and Developers.

Flow from AI Development to Utilization (Reference)

AISI Japan AI Safety Institute

AI Developer (Foundational model development companies)	AI Provider (Software companies or cloud service companies that integrate AI models into proprietary systems)	AI User
Provides APIs to make developed AI functionalities accessible to other organisations. (May also release AI functionalities as open-source.)	Can focus on developing in-house services by utilizing pre-trained models through APIs, without needing to develop AI models themselves.	Utilize AI service



AI Developer	AI capabilities (API)	Key Features
OpenAI	ChatGPT API	Generates text in conversational format.
Google	Google Cloud Natural Language API	Analyses sentiment in text.
Stability AI	Stable Diffusion API	Generates images from text.

Examples of AI-Driven Services/Systems:
• Customer support chatbots.
• Tools for analyzing and categorizing review site posts.
• Character design tools.

End Users (Use Cases):
Customers: Asking questions through chatbots about products.
Employees: Using review analysis results for marketing.
Employees: Exploring character designs for advertising.

The Importance of AI Safety Measures for AI Providers

AISI Japan AI Safety Institute

- As providing AI services becomes easier without the need to develop models from scratch, addressing AI-specific risks has become crucial for service providers, alongside the usual quality management of applications.
- Insufficient safeguards may lead to compliance violations, reputational damage, revenue loss, legal claims, or business suspension.

Significant AI Risk Incidents with Potential Major Impact

#	Overview	Impact(Example)
1	An AI driven hiring system favoured male candidates, disadvantaging female applicants.	Female applicants lost job opportunities.
2	A lawyer used a generative AI chat tool for legal research, but it produced false information. The lawyer submitted a document with fabricated details to the court and was fined for this misconduct.	The user (lawyer) faced penalties (fines).
3	A generative AI chatbot gave inappropriate advice to a male user, reportedly contributing to his while he was in a vulnerable state.	Loss of human life.
4	A system performing image recognition mistakenly identified a human as a gorilla, causing a major issue.	Significant dagame to the company's credibility and brand image.

AI safety measures are essential for providers, even when offering products developed using AI models from external companies

AISI Japan AI Safety Institute

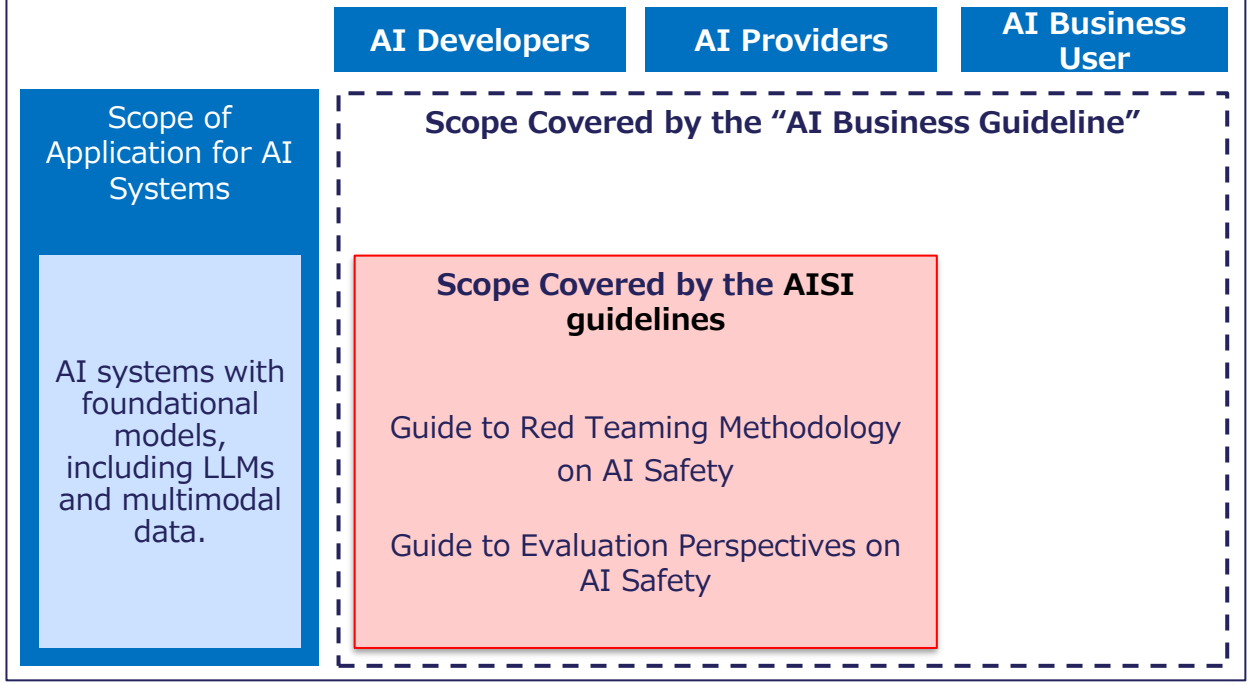
Introduction to Key Guidelines for Realizing AI Safety

- Businesses involved in the development, provision, and use of AI must refer to the AI business guidelines and consider appropriate measures.
- For developers and providers of AI systems using foundational models (including large-scale language models, LLMs) or handling multimodal information, referencing AISI's guidelines can effectively support AI safety evaluation and testing.

Key guidelines Related to AI Safety

Publisher	Title	Overview
Ministry of Economy, Trade and Industry	AI Business Guideline	A unified set of national recommendations for promoting safe and secure use of AI in business operations.
AISI	Guide to Red Teaming Methodology on AI Safety	A basic framework for considerations when conducting AI safety evaluations.
AISI	Guide to Evaluation Perspectives on AI Safety	A foundational guide for assessing risks in AI systems, including attack vectors and scenarios, through red-teaming methodologies.

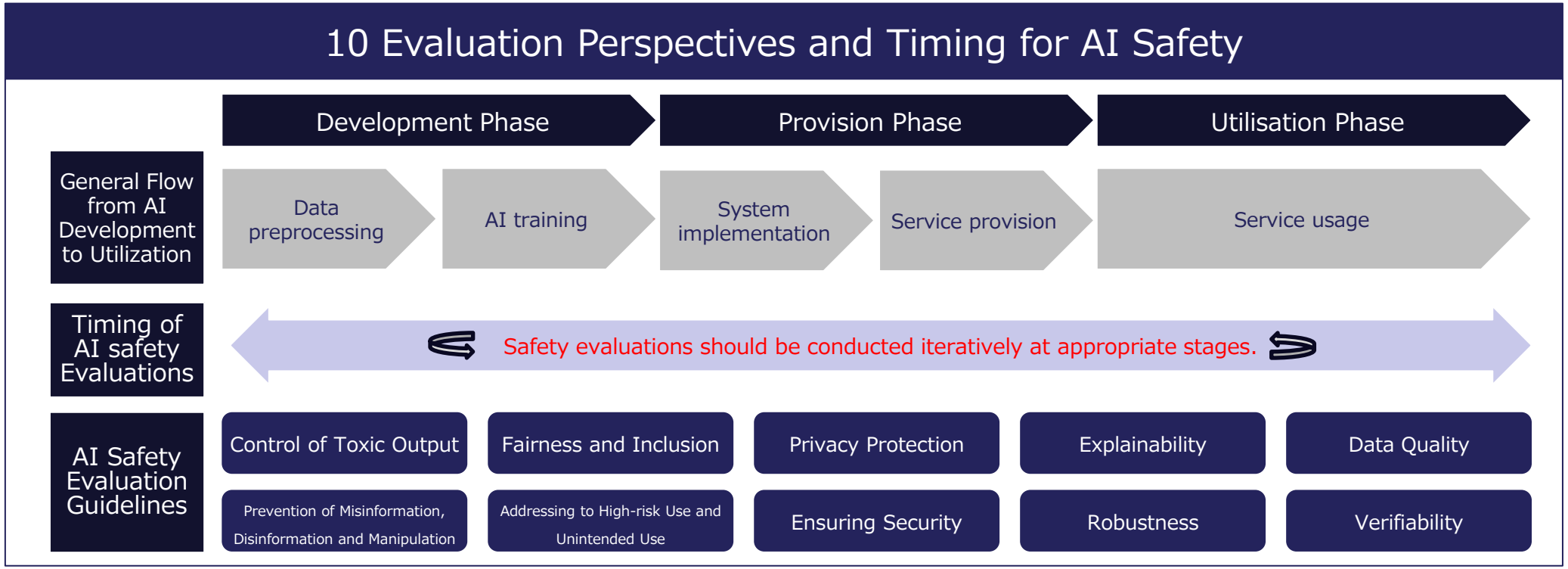
Intended Audience of Each Guideline



AISI Japan AI Safety Institute

Implementing Risk Measures Aligned with Organizational Scale and Resources

- The AI Safety Evaluation Perspective Guide defines 10 key evaluation perspectives for AI safety.
- It recommends conducting safety evaluations and implementing risk measures for AI systems and services.
- Prioritizing measures for services with higher risk tolerance or significant impact is effective.
- AI safety evaluations should be conducted not only during development and provision but also regularly after service launch to ensure ongoing safety.



Summary for guidelines in AI safety

AISI Japan AI Safety Institute

Overview of Evaluation Perspectives

Please refer to the “Evaluation Perspectives on AI Safety” below and verify whether appropriate measures have been taken for each identified risk.

Evaluation Perspectives on AI Safety		Overview of Evaluation Perspectives	Evaluation Methods (Examples)
1	Control of Toxic Output	Ensuring the appropriateness of AI-generated outputs by filtering out inappropriate information	• Does the outputs contain violent expressions or information that could cause psychological harm? • Does the outputs aids in committing crimes or promotes illegal activities?
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	Verifying the basis of AI-generated outputs to ensure end users can use AI safely	• Does the AI service output display source information, making the basis of the output visible?
8	Robustness		
9	Data Quality	Ensuring the appropriateness of data accessed by AI	• Does the data include malicious or faulty programs? • Does the data include personal information or copyrighted materials?
10	Verifiability		

AISI Japan AI Safety Institute

Benefits of Utilizing Guidelines for AI Safety

- Efficiently addressing AI-related risks ensures the delivery of reliable and trustworthy services.
- Adhering to the guidelines enhances an organization's credibility both internally and externally.

Please also refer to the "AI Guidelines for Business," which served as a reference in developing the AISI Guidelines. These guidelines help AI service providers in Japan use AI appropriately

AISI Japan

Download guidelines here

Guide to Evaluation Perspectives on AI Safety

Guide to Red Teaming Methodology on AI Safety

AI Guidelines for Business

AISI

Japan AI Safety Institute