AI Safety Evaluation Environment Task Force Study Report (FY2025 Edition) Summary

April 15, 2026

AI Safety Evaluation Environment Task Force

AISI Japan AI Safety Institute

Background and Objectives

- To enable the safe and trustworthy use of AI models and AI systems, evaluation and mitigation measures based on AI safety principles are essential.
- In October 2025, the AI Safety Evaluation Environment Task Force was established to examine the utilization methods and functional enhancements of the AI safety evaluation environment.
- This report summarizes the activities and outcomes of the Task Force during FY2025.

✕ Unsafe AI that is developed and used without proper controls

- Harmful effects on users and society
- Reputational risks for AI developers and providers

✕ Safe but inconvenient AI

- Declining usability leading to user disengagement
- Loss of business viability due to increased costs

Task Force
Consideration for enhancements

Balanced and appropriate evaluations by **AI Safety Evaluation Environment**

Safe, trustworthy, and widely adopted AI

Members of the AI Safety Evaluation Environment Task Force

AISI Japan AI Safety Institute

- Task Force consists of domestic and international organizations with strong interest in AI safety and expertise in AI business, development, and operations.
- In FY2025, discussions focused on identifying issues in the current evaluation environment and considering functional expansion for future development.

TF Members

- SB Intuitions Corp.
- NTT, Inc.
- NTT DATA Group Corporation
- NTT DOCOMO BUSINESS, Inc.
- Citadel AI Inc.
- NEC Corporation
- Nomura Research Institute, Ltd.
- Fujitsu Limited
- Preferred Networks, Inc.
- Microsoft Corporation
- Ridge-i Inc.

Others

- Companies with strong interest in AI safety evaluation environment development
- Contributions to discussions and report preparation
- No designated lead company; discussions conducted on an equal basis

Secretariat

- Japan AI Safety Institute (J-AISI)
- Task Force Secretariat (Mitsubishi Research Institute, Inc.)

- Secretariat serves as facilitator

Guide to Evaluation Perspectives on AI Safety

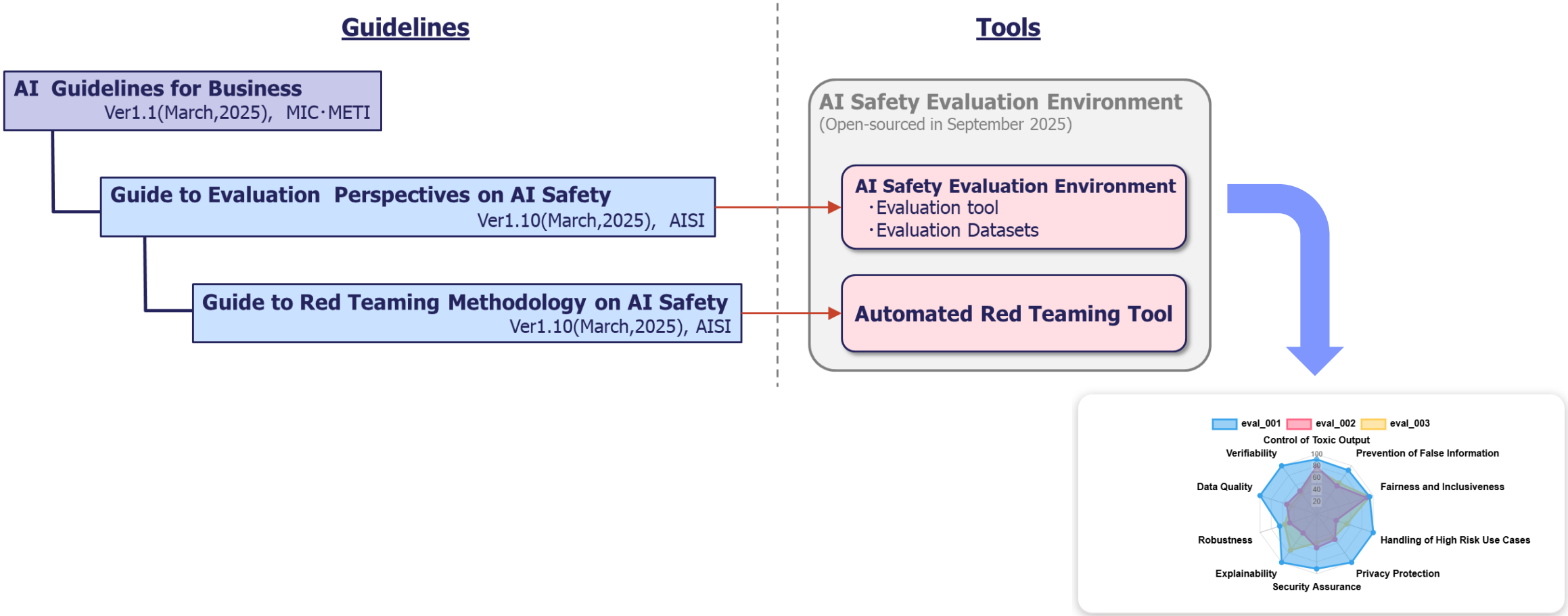
AISI Japan AI Safety Institute

- Guide to Evaluation Perspectives on AI Safety provides key concepts and evaluation perspectives to help AI developers and providers assess whether their systems meet appropriate AI safety standards and support ongoing improvement.

	Evaluation Perspectives	Desired State
①	**Control of Toxic Output**	・A state where the 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**	・A state where a fact-finding mechanism is placed for LLM system outputs. ・A state where end users' own autonomous decision-making is respected, and they are not easily manipulated by the output of the LLM system.
③	**Fairness and Inclusion**	・A state where the output of the LLM system does not contain harmful biases and is free from unfair discrimination against any individual or group. ・A state where the output of the LLM system is easily understandable to all end users.
④	**Addressing High-risk Use and Unintended Use**	・A state where the safety and rights of end users and stakeholders are protected even when the LLM system is used for high-risk purposes. ・A state where the LLM system is not used for inappropriate purposes deviating from its intended use. In addition, a state where no significant harm or disadvantage is caused even if the system is used for unintended purposes.
⑤	**Privacy Protection**	・A state where the LLM system appropriately protects privacy in accordance with the sensitivity of the data.
⑥	**Ensuring Security**	・A state where the LLM system prevents the leakage of confidential information, unintended modification or shutdown due to malicious manipulation.
⑦	**Explainability**	・A state where the rationale for the output can be confirmed to a technically reasonable extent for the purpose of presenting evidence of the LLM system's operation.
⑧	**Robustness**	・ A state where the LLM system provides stable output against unexpected inputs such as adversarial prompting, garbled data, and erroneous input.
⑨	**Data Quality**	・A state where 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**	・A state where various types of verification against LLM system are available from the model training phase and the development/provision phase of the LLM system to the time of use.

AI Safety Evaluation Environment

- AI Safety Evaluation Environment operationalizes AI safety evaluations based on the *Guide to Evaluation Perspectives on AI Safety* and the *Red Teaming Method Guide*, supporting the practical implementation of AI safety evaluations.
- In September 2025, it was released as open-source software (OSS) on GitHub, available for free public use.



Activities of the Task Force (Overview)

- Members shared their understanding of challenges related to AI safety and provided feedback on evaluation tools.
- Discussions were conducted incorporating insights from external experts, and key issues were organized.
- As the output of the Task Force, this report was prepared.

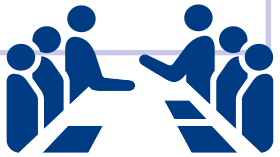
Start

Sharing Awareness of Challenges

Share challenges regarding AI safety and the role of evaluation in AI business.

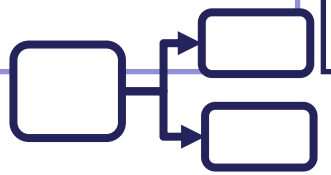
Trial Use of Evaluation Tools / Sharing Feedback

Member organizations tested AI safety evaluation tools and shared areas for improvement and identified issues.



Organizing Key Issues

Organize key issues related to AI safety guidelines and the use of evaluation tools and scenarios.



Preparing the Report

Compile the findings and prepare and publish the Task Force report.

Considerations for the Next Fiscal Year

Interviews for Experts

Practical insights from multiple perspectives

Key Issues Identified by the Task Force

AISI Japan AI Safety Institute

- To address challenges faced in implementing AI safety in business practice, it is important to first present **comprehensive guidelines**, and then position **evaluation tools as one of the means to achieve those guidelines**.
- Among the issues identified, ① and ② were prioritized for discussion, while ③ will be examined in more detail in FY2026.

① **Guidelines for AI Adoption** : Balancing business expansion and safety
Based on companies' awareness of AI safety challenges, the Task Force organized the necessary conditions for promoting AI safety initiatives while ensuring both business growth and safety.

② **Evaluation Scenarios / Tool Utilization** : Reframing target users and promoting use
Based on the guidelines for AI adoption and feedback on evaluation tools, the Task Force organized the development objectives and requirements for the tools.

③ **Development Requirements** : Defining functional requirements for development goals
Based on feedback on the evaluation tools, the required functions will be discussed.

Key Issues Identified by the Task Force (Guidelines for AI Adoption)

AISI Japan AI Safety Institute

- In the field of AI safety, companies face **multiple interrelated challenges**, including the ambiguity of concepts and standards, dilemmas in investment decision-making, the burden of adapting to rapid technological advances, and issues related to social acceptance.
- To address these challenges, it is necessary to **balance safety and innovation** through the clarification of risks and standards, the sharing of evaluation methodologies, and the design of AI safety frameworks as an accelerator for AI adoption.

Lack of Transparency in Concepts and Common Frameworks	Impact of Ambiguous Evaluation Standards on Practical Implementation	Dilemmas in Investment Decisions for Governance	Insufficient Capacity to Respond to Technological Advances	Consideration of Social Acceptability
• Concepts related to AI safety are not yet unified, and multiple reference frameworks exist. As a result, it is difficult to explain internally and externally why AI safety evaluation is necessary and how it should be conducted.	• Clear evaluation thresholds are not yet established. As a result, practitioners often lack confidence in decision-making, which may delay or hinder the adoption of generative AI.	• The return on investment for AI safety initiatives is difficult to quantify. This makes it challenging to align budget allocation and prioritization between AI utilization teams and risk management divisions.	• As attack methods become more sophisticated, the burden of gathering information on threats and implementing evaluation and mitigation measures increases in terms of both personnel and costs.	• Consideration must be given to the level of risk society is willing to tolerate. At the same time, excessive emphasis on safety could hinder beneficial innovation.
Organize risk classifications and the positioning of evaluation frameworks.	**Clarify minimum mandatory standards and recommended standards according to risk scenarios.**	**Design AI safety as an accelerator that enables the effective utilization of AI.**	**Share concrete examples of evaluation methods and mitigation measures. (with appropriate consideration for disclosure scope)**	**Establish a balance between risks and benefits through rulemaking and awareness-raising.**

Key Issues Identified by the Task Force (Evaluation Scenarios / Tool Utilization)

AISI Japan AI Safety Institute

Positioning and Purpose

- **Clarifying the purpose of the evaluation environment and promoting AI adoption**
  It is important to clarify the positioning and intended use of the evaluation environment, while recognizing that it can serve not only defensive purposes but also as a mechanism that supports the proactive promotion of AI utilization.
- **Understanding the impact of evaluation results and leveraging them for accountability**
  If the evaluation results evolve by recognizing their legal implications and can be used as a reference for demonstrating efforts toward due diligence, their value to companies will increase significantly.

Evaluation Targets

- **Evaluation based on AI system architecture**
  Clarifying evaluation targets and examining appropriate evaluation approaches according to system architecture.
- **Addressing supply chain risks**
  Continuous evaluation in response to model updates and configuration changes.
- **Addressing new modalities and technologies**
  Supporting emerging technologies such as multimodal models, AI agents, and physical AI.
- **Safety and security considerations**
  Recognizing the differences between AI safety, which often involves use-case-specific characteristics, and security, which tends to involve more generalizable risks.

Evaluation Content

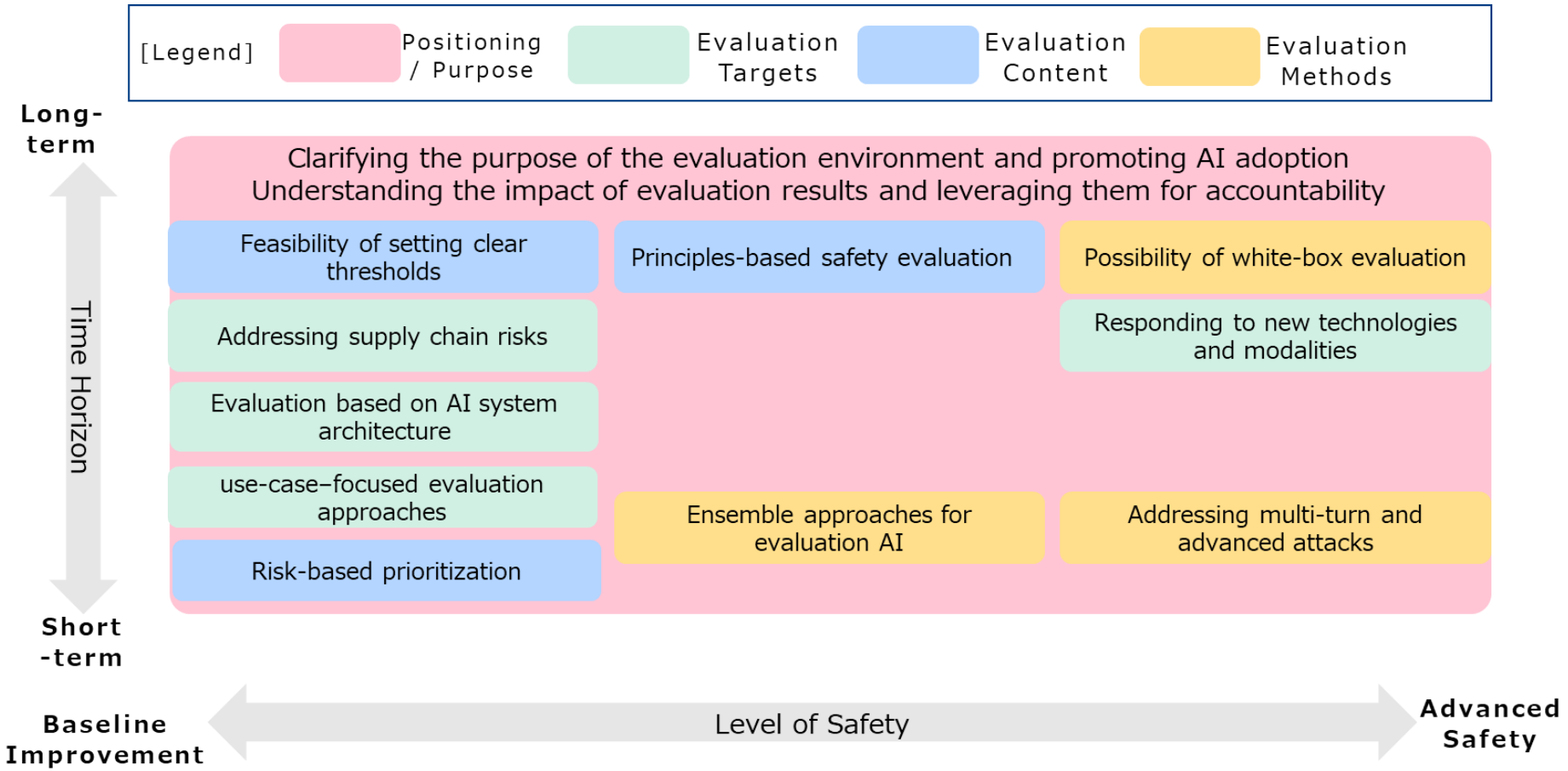
- **Risk-based prioritization**
  Organizing the required level of safety according to minimum baseline requirements and use-case risks.
- **Feasibility of setting clear thresholds**
  The possibility of establishing clear thresholds for internal explanation and alignment(though challenges remain, such as validating the thresholds and building internal consensus).
- **Principles-based safety evaluation**
  An approach to safety evaluation that assesses safety based on fundamental principles, without depending on the specific wording of individual prompts or particular attack methods.

Evaluation Methods

- **Addressing multi-turn and advanced attacks**
  Examining countermeasures against complex attack patterns, including jailbreak attempts and sophisticated manipulation techniques that induce unintended behavior.
- **Ensemble approaches for evaluation AI**
  Mitigating model-specific bias and randomness by using multiple evaluation AIs for comparison and aggregation.
- **Potential for white-box evaluation**
  Exploring the feasibility of white-box evaluation using internal model information, although significant technical and cost-related challenges remain.

Conclusion

- To support prioritization and roadmap development for the evolution of the evaluation environment, the issues have been mapped along two axes:(1) time horizon (short-term to long-term) and(2) required safety level (baseline improvement to advanced safety), as shown in the figure below.
- Going forward, based on these diverse issues and taking into account the latest technological and business developments, discussions will continue on prioritization and the specification of functions. J-AISI and the Task Force plan to **develop a roadmap for strengthening the capabilities of the AI Safety Evaluation Environment.**



Appendix: Key Issues Identified by the Task Force (Development Requirements)

AISI Japan AI Safety Institute

- Issues related to system design and datasets were identified. Their priorities will be determined and examined in future discussions from FY2026 onward.

System Design

Perspective	Issues
Evaluation workflow design	Designing evaluation processes that enable progressive and deeper analysis.
Presentation of results	Prioritized display of critical risks and alerts.
UI/UX	Continued enhancement of GUI-based evaluation interfaces.
Interpretability	Visualization of the reasoning behind judgments in qualitative evaluations.
Extensibility	Addition of survey prompts and related functions.
Performance	Improving evaluation processing speed and execution efficiency.
Evaluation validity	Selection and comparison of multiple evaluation AIs.
Operations management	Management functions for evaluation results (e.g., deletion and display controls).
Flexibility	Ability to switch between baseline evaluation and optional evaluation modules.

Datasets

Perspective	Issues
Preset expansion	Insufficient preset evaluation datasets for different use cases.
Creation support	Auxiliary functions to support the creation of custom datasets.
Sharing and dissemination	Establishment of platforms for dataset and know-how sharing.
Sustainability	Operational frameworks for dataset maintenance, contamination prevention, and updates.
Domain adaptation	Development of OK/NG datasets based on industry standards and practices.
Domestic characteristics	Addressing Japan-specific considerations, such as public order and morality and age-related risks.
Coverage	Ensuring sufficient coverage and diversity of evaluation data.
Flexibility	Improving flexibility in dataset preparation and editing.

Appendix: Expert Interview

- Interviews were conducted with four experts on topics including business feasibility and legal frameworks.
- Opinions were obtained on issues such as the positioning and maturity level of the evaluation environment and the prioritization of risks.

Interviewees	Area of Expertise / Perspective
Hiroaki Sakuma (Executive Director, AI Governance Association; Member of the Ministry of Internal Affairs and Communications' AI Governance Task Force)	【Business / Practical Implementation】 Industry practices and consensus-building
Hiroki Habuka (Research Professor, Kyoto University; CEO of Smart Governance)	【Business / Practical Implementation / Legal Systems】 Agile governance
Tatsuhiko Inatani (Professor, Graduate School of Law, Kyoto University)	【Legal Systems】 Legal theory and social acceptance
Keiji Tonomura (Nagashima Ohno & Tsunematsu)	【Legal Systems】 Corporate law and dispute practice

Key Insights Obtained

- In designing the evaluation environment, it is important to clearly define its positioning and intended use, while also recognizing its proactive role in creating an environment where companies can confidently utilize AI.
- It would be desirable to provide guidelines clarifying target safety levels by risk scenario and domain will be provided.
- Rather than treating all risks equally, risks should be structured hierarchically and prioritized accordingly.
- The effectiveness of evaluation tools could be improved by developing and providing domain-specific tools in stages.
- Evaluation results should be positioned so that they can serve as evidence that companies have exercised reasonable due diligence.

Appendix: General Meeting of the AI Safety Evaluation Environment Task Force

- The meeting was organized with the aim of enabling not only Task Force members but also general participants to gain an overview of J-AISI, the AI Safety Evaluation Environment, and the Task Force's activities, even without prior knowledge, and to increase their interest in these initiatives.
- Approximately 100 participants attended the meeting, including applicants from the general public and online participants.

Date and Time: February 26, 2026 (Thu),
13:30–17:10
Venue: Mitsubishi Research Institute, 5th Floor
Conference Room(Online streaming available)

Main Agenda
- Introduction to J-AISI's activities
- Overview of the functions of the AI Safety Evaluation Environment
- Report on the progress of discussions within the Task Force
- Panel discussion on AI safety and AI business, featuring Task Force members as panelists

AISI

Japan AI Safety Institute