Vision Paper

# Business Demonstration Project to Promote AI Utilization through AI Safety Evaluation

## — Toward a Society Where AI Trustworthiness and Innovation Coexist—

July 10, 2025

Japan AI Safety Institute
Business Demonstration Working Group

AISI Japan
AI Safety Institute

# Purpose of This Document

- This Vision Paper aims to serve as both an agenda and a guideline for realizing an AI society in which trustworthiness and innovation can coexist. It presents a common understanding and a pathway toward the concrete implementation of AI safety evaluation at the levels of society, industry, and policy.

**Definition of "AI Safety"**

*"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. (*Social risks include physical, psychological, and economic risks.)"*

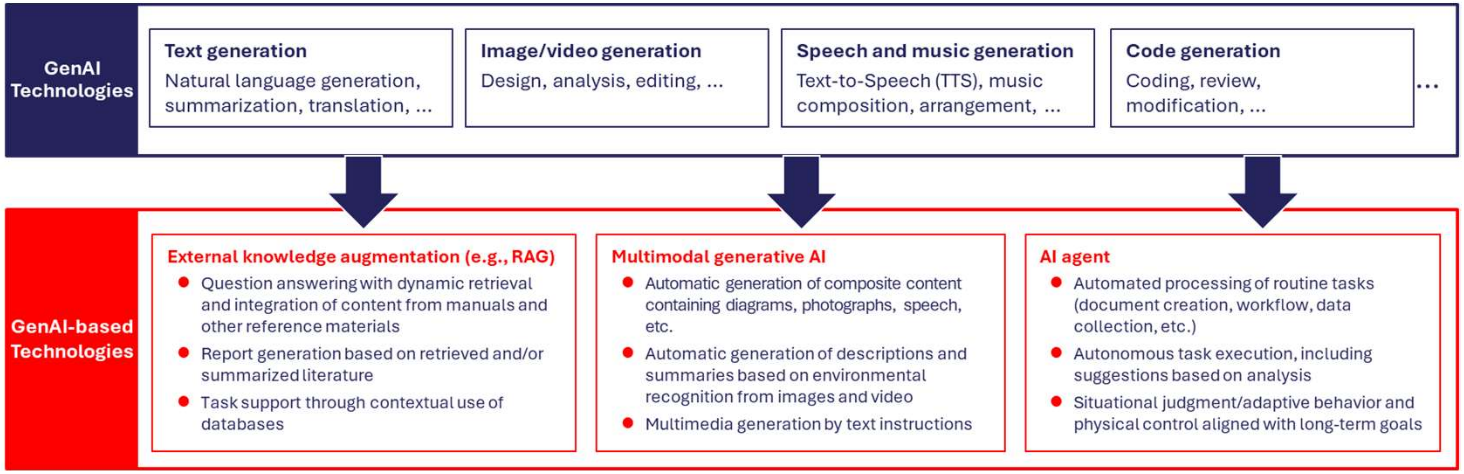
AISI, Guide to Evaluation Perspectives on AI Safety

**Target audience**

| AI developers and providers | Companies and Municipalities Implementing and Utilizing AI | Policymakers and certification bodies |
|---|---|---|
| To be used as a reference when conducting AI safety evaluations of foundation models, products, and services. | To serve as a guide for identifying and organizing issues related to AI safety and understanding the factors involved in its social acceptability. | To inform future discussions on the development of AI safety evaluation guidelines. |

# Current Landscape of AI Utilization

- The advancement of generative AI (GenAI)-based technologies is bringing about qualitative changes in the societal role of AI and its positioning within systems.



**GenAI Technologies**

**Text generation**
Natural language generation, summarization, translation, ...

**Image/video generation**
Design, analysis, editing, ...

**Speech and music generation**
Text-to-Speech (TTS), music composition, arrangement, ...

**Code generation**
Coding, review, modification, ...

...

**GenAI-based Technologies**

**External knowledge augmentation (e.g., RAG)**

- Question answering with dynamic retrieval and integration of content from manuals and other reference materials
- Report generation based on retrieved and/or summarized literature
- Task support through contextual use of databases

**Multimodal generative AI**

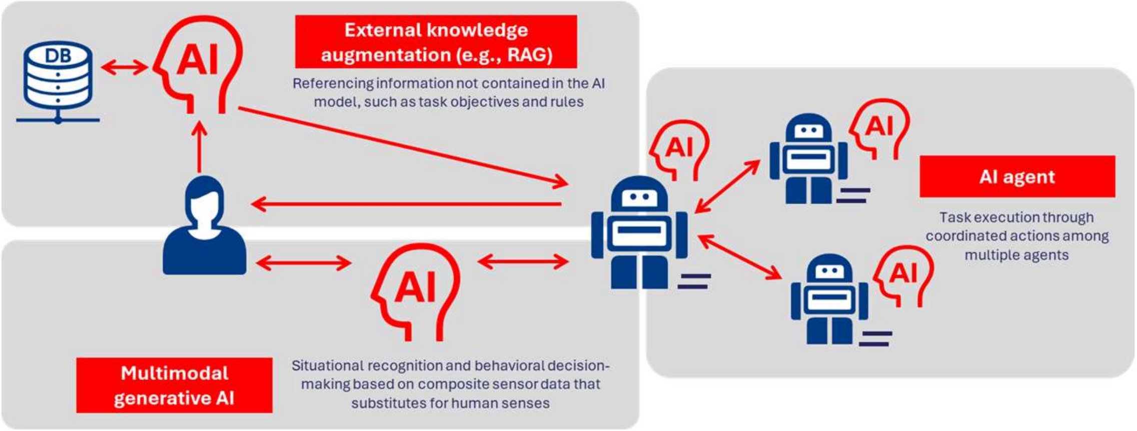
- Automatic generation of composite content containing diagrams, photographs, speech, etc.
- Automatic generation of descriptions and summaries based on environmental recognition from images and video
- Multimedia generation by text instructions

**AI agent**

- Automated processing of routine tasks (document creation, workflow, data collection, etc.)
- Autonomous task execution, including suggestions based on analysis
- Situational judgment/adaptive behavior and physical control aligned with long-term goals

GenAI helps automate and support business processes through text generation, summarization, translation, and classification.

Interest is growing in GenAI-based technologies for advanced processing and decision support.



**External knowledge augmentation (e.g., RAG)**
Referencing information not contained in the AI model, such as task objectives and rules

**AI agent**
Task execution through coordinated actions among multiple agents

**Multimodal generative AI**
Situational recognition and behavioral decision-making based on composite sensor data that substitutes for human senses

Beyond digital tasks, AI is entering the physical world, supporting behavior control and operations.

Examples include:

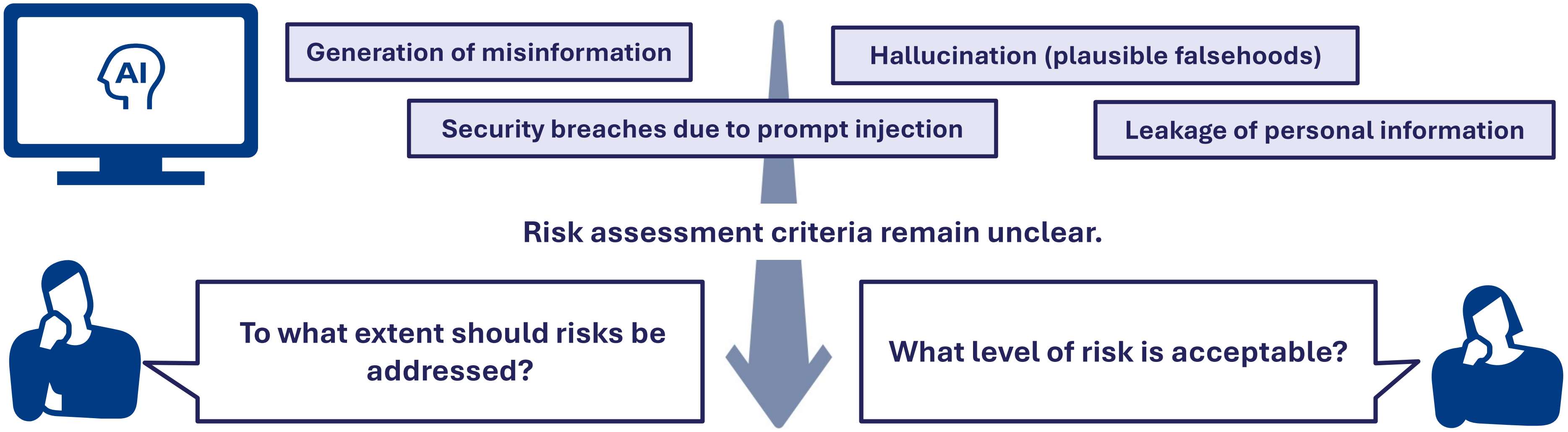
- Drones for disaster search
- Robots for coordinated transport in factories

**A new phase of AI as social infrastructure**

# Necessity of AI Safety Evaluation

- As the utilization of AI rapidly expands and brings new benefits, numerous risk events that could undermine public trust have also begun to surface.
- There is an urgent need to develop a framework for AI safety evaluation that considers the application and risk characteristics of AI.

**Emerging risks associated with the utilization of AI**



AI
Generation of misinformation
Hallucination (plausible falsehoods)
Security breaches due to prompt injection
Leakage of personal information
Risk assessment criteria remain unclear.
To what extent should risks be addressed?
What level of risk is acceptable?

**Urgent need for a framework for AI safety evaluation considering AI's application and risk characteristics**

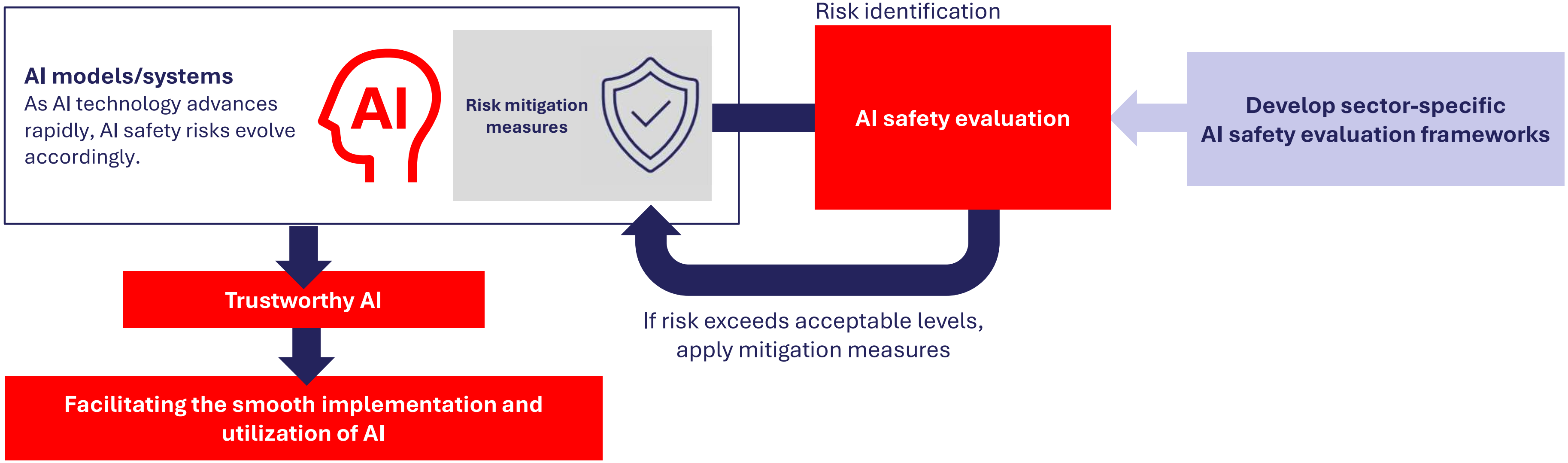
Appropriate evaluation of AI safety is essential in achieving both trustworthiness and innovation
Demonstrating that AI is a trustworthy technology then provides a foundation for AI to be accepted by society as a "trustworthy AI.".

# Objectives and Goals of the Business Demonstration WG

AISI Japan AI Safety Institute

- The Business Demonstration WG aims to develop a framework for AI safety evaluation—based on users' understanding of risks—by building mechanisms to support the assurance of AI safety in the societal implementation of AI through collaboration among industry, government, and experts.

| Medium- to long-term goals | Through the development of an AI safety evaluation framework:<br>(1) to create opportunities that contribute to solving societal challenges—such as those related to healthcare, labor, and aging—by promoting the smooth introduction and widespread adoption of AI across industries, thereby supporting its societal implementation; and<br>(2) to develop an environment in which evaluation methods and perspectives function as a common language that is easy to understand and apply for both AI developers/providers and users. |
|---|---|

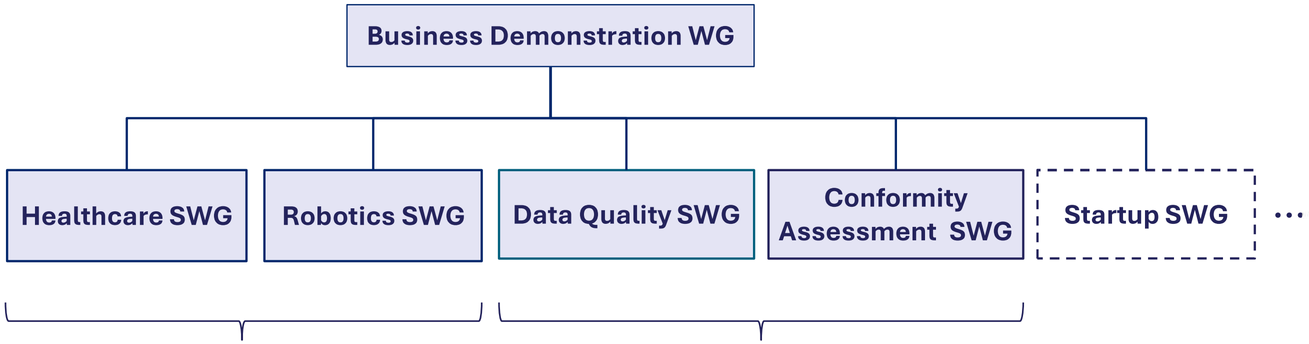


**AI models/systems**
As AI technology advances rapidly, AI safety risks evolve accordingly.

AI
Risk mitigation measures
Risk identification
AI safety evaluation
Develop sector-specific AI safety evaluation frameworks
If risk exceeds acceptable levels, apply mitigation measures
Trustworthy AI
Facilitating the smooth implementation and utilization of AI

# Structure of the Business Demonstration WG

- The Business Demonstration WG combines sector-specific (Vertical) SWGs and cross-sectoral (Horizontal) SWGs.
- In FY2025, it consists of four SWGs: Healthcare, Robotics, Data Quality, and Conformity Assessment.

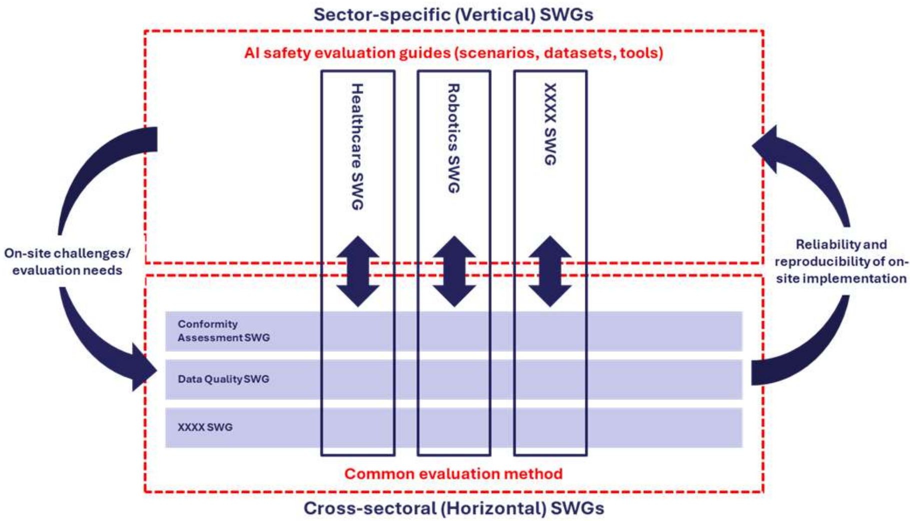


**Sector-specific (Vertical) SWG:**
**Develop sector-specific evaluation guides**

- **Sectors of high international interest where Japan has strengths**
- **Areas where the advancement of AI utilization and ensuring AI safety is especially important**

**Cross-sectoral (Horizontal) SWG:**
**Develop cross-sectoral evaluation methods**

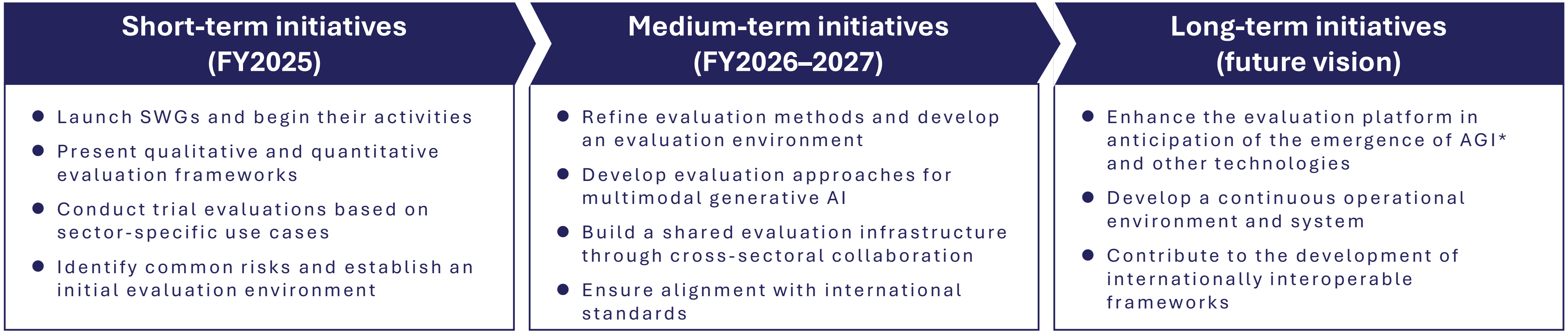
- **Building multilayered evaluation frameworks based on Japan-originated practices**
- **Contributing to the development of international and standardized rules**


Sector-specific (Vertical) SWGs
AI safety evaluation guides (scenarios, datasets, tools)
Healthcare SWG
Robotics SWG
XXXX SWG
On-site challenges/ evaluation needs
Reliability and reproducibility of on-site implementation
Conformity Assessment SWG
Data Quality SWG
XXXX SWG
Common evaluation method
Cross-sectoral (Horizontal) SWGs

- **Each SWG organizes evaluation perspectives, develops guidelines and tools, and coordinates with other SWGs. Additional SWGs will be established as needed.**
- **Vertical and Horizontal SWGs play their respective roles while complementing one another and working together.**

# Roadmap of the Business Demonstration WG

- To promote the societal implementation of AI, the Business Demonstration WG will take a phased and continuous approach toward establishing AI safety evaluation.
- In FY2025, the WG will form an initial evaluation foundation through use cases in key areas.
- In the medium to long term, it will develop evaluation methods and environments that accommodate technological diversity, while promoting alignment with international standards.



**Short-term initiatives (FY2025)**

- Launch SWGs and begin their activities
- Present qualitative and quantitative evaluation frameworks
- Conduct trial evaluations based on sector-specific use cases
- Identify common risks and establish an initial evaluation environment

**Medium-term initiatives (FY2026–2027)**

- Refine evaluation methods and develop an evaluation environment
- Develop evaluation approaches for multimodal generative AI
- Build a shared evaluation infrastructure through cross-sectoral collaboration
- Ensure alignment with international standards

**Long-term initiatives (future vision)**

- Enhance the evaluation platform in anticipation of the emergence of AGI* and other technologies
- Develop a continuous operational environment and system
- Contribute to the development of internationally interoperable frameworks

*Artificial General Intelligence

# Healthcare SWG

- While generative AI is increasingly used to reduce the burden on healthcare professionals and support health management for general consumers, concerns over information handling and incorrect outputs are prompting the development of utilization guidelines.
- There is a pressing need to establish an evaluation framework that can predict, control, and verify the potential impact of generative AI outputs on users.

**Current Landscape of AI Utilization in Healthcare**

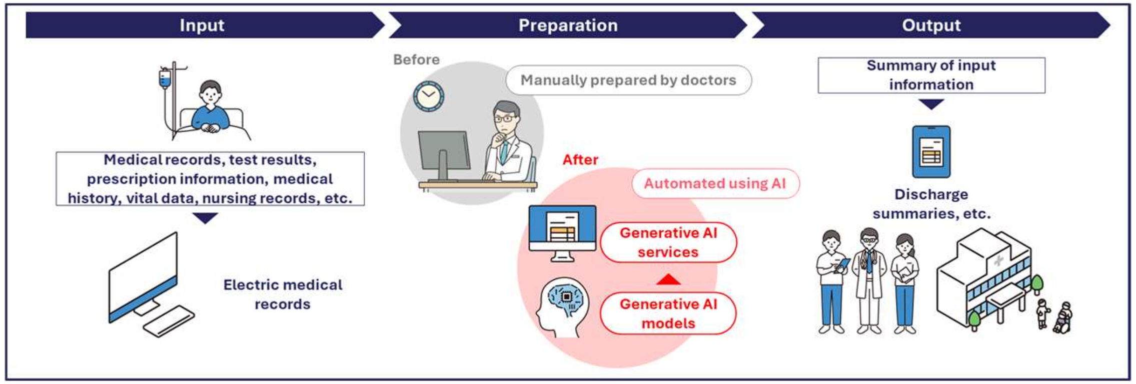
- Regulations on overtime work at medical institutions
  - ➢ Reducing administrative burden such as documentation by healthcare professionals
- Widespread availability of products and services for personal health management
- Sector-specific initiative: “Guide to the Utilization of Generative AI for Healthcare Providers”*

**AI Safety Evaluation: Importance & Expectations for Utilization**

- The healthcare sector involves potential risks to life
  - ➢ Concerns about hallucinations and information handling
- JaDHA guidelines and criteria provide some visibility
  - ➢ Evaluation items and methods remain insufficient
- Need to systematize use-case-based evaluations
  - ➢ To ensure a safe and secure utilization environment

*Japan Digital Health Alliance (JaDHA) ,“Guide to the Utilization of Generative AI for Healthcare Providers” (Version 2.0, Feb. 2025)

**Examples of generative AI utilization in healthcare（discharge summary preparation, etc.）**



Input

Medical records, test results, prescription information, medical history, vital data, nursing records, etc.

Electric medical records

Preparation

Before

Manually prepared by doctors

After

Automated using AI

Generative AI services

Generative AI models

Output

Summary of input information

Discharge summaries, etc.

Develop an AI safety evaluation framework based on practical needs and risk concerns, referencing existing guidelines.

# Healthcare SWG

- Utilizing generative AI in healthcare involves various risks. It is essential to develop use-case-specific evaluation perspectives and both quantitative & qualitative evaluation models.
- The healthcare SWG aims to enhance the practical effectiveness of AI safety evaluations through the development of checklists, evaluation tools, and feasibility assessments using actual data.

**Potential Risks**

- Privacy leakage
  - ➢ Incomplete masking of data
- Hallucination
  - ➢ Medical errors or the provision of incorrect information
- Lack of explainability
  - ➢ Absence of transparent reasoning behind outputs can hinder understanding in medical settings

**Direction of AI Safety Evaluation**

- Develop evaluation perspectives for each use case
- Build both quantitative & qualitative evaluation models
- Target text-generating AI classified as Non-SaMD*
- Enhance the practical effectiveness of AI safety evaluation through:
  - Development of checklist-style evaluation item sets
  - Development of evaluation tools
  - Feasibility testing using actual data

**SWG Roadmap**

**Short-term initiatives (FY2025)**

- Select representative use cases subject to consideration with a focus on Non-SaMDs
- Clarify risk structure associated with generative AI utilization and design AI safety evaluation perspectives and scenarios
- Consider evaluation guides, datasets, and tools

**Medium-term initiatives (FY2026–2027)**

- Support the development and verification of evaluation guides, datasets, and tools by AI developers and providers

  Example: Conduct pilot AI safety evaluations of generative AI products and services implemented in medical institutions

**Long-term initiatives (future vision)**

- Engage in promotion and dissemination activities to encourage the use of guides, datasets, and tools within the healthcare sector
- Continuously update evaluation guides, datasets, and tools

*Non-SaMD refers to programs primarily aimed at promoting health and other similar purposes, which are not subject to regulation under Japan's Pharmaceuticals and Medical Devices Act.

# Robotics SWG

- In robotics, AI safety needs to be considered based on human-robot communication and collaboration.
- This requires building an evaluation framework that balances the benefits of AI-robot integration with the inherent risks involved.

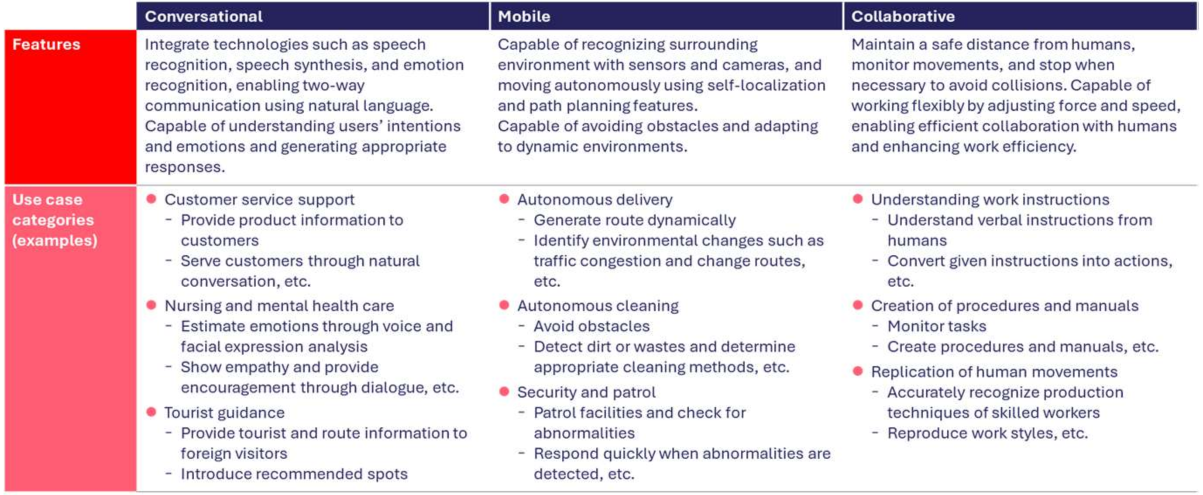
**Current Landscape of AI Utilization in Robotics**

- Exploring the potential of multimodal AI in human-interface applications to replicate human decision-making based on the five senses
- Human-robot and robot-robot collaboration in cooperative robotic systems
- Advanced integration of situational understanding and action selection driven by progress in embodied AI

**AI Safety Evaluation: Importance & Expectations for Utilization**

- Human-interaction risks and situational-response challenges in conversational, mobile, and collaborative robots
  - ➢ These cannot be fully addressed by functional safety
- To cover use cases such as cooperative behavior, a framework is needed to classify and evaluate robots using a matrix of "robot types × AI functions"

## Designing evaluation methods based on scenarios developed from specific use cases



| | Conversational | Mobile | Collaborative |
|---|---|---|---|
| Features | Integrate technologies such as speech recognition, speech synthesis, and emotion recognition, enabling two-way communication using natural language. Capable of understanding users' intentions and emotions and generating appropriate responses. | Capable of recognizing surrounding environment with sensors and cameras, and moving autonomously using self-localization and path planning features. Capable of avoiding obstacles and adapting to dynamic environments. | Maintain a safe distance from humans, monitor movements, and stop when necessary to avoid collisions. Capable of working flexibly by adjusting force and speed, enabling efficient collaboration with humans and enhancing work efficiency. |
| Use case categories (examples) | ● Customer service support<br>- Provide product information to customers<br>- Serve customers through natural conversation, etc.<br>● Nursing and mental health care<br>- Estimate emotions through voice and facial expression analysis<br>- Show empathy and provide encouragement through dialogue, etc.<br>● Tourist guidance<br>- Provide tourist and route information to foreign visitors<br>- Introduce recommended spots | ● Autonomous delivery<br>- Generate route dynamically<br>- Identify environmental changes such as traffic congestion and change routes, etc.<br>● Autonomous cleaning<br>- Avoid obstacles<br>- Detect dirt or wastes and determine appropriate cleaning methods, etc.<br>● Security and patrol<br>- Patrol facilities and check for abnormalities<br>- Respond quickly when abnormalities are detected, etc. | ● Understanding work instructions<br>- Understand verbal instructions from humans<br>- Convert given instructions into actions, etc.<br>● Creation of procedures and manuals<br>- Monitor tasks<br>- Create procedures and manuals, etc.<br>● Replication of human movements<br>- Accurately recognize production techniques of skilled workers<br>- Reproduce work styles, etc. |

A new AI safety evaluation framework is being developed beyond functional safety, including trust-building and context understanding.

# Robotics SWG

- ■ When conducting AI safety evaluations in the robotics sector, it is essential not only to address reduction of physical risks but also to identify and control psychological and societal risks.
- ■ The Robotics SWG will promote multilayered evaluations by robot type using both real-world settings and simulated scenarios, with the aim of establishing a future common framework.

**Potential Risks**

- Physical risks
  - ➢ Collision, contact, electric shock, fire, etc.
- Psychological risks
  - ➢ Fear, intimidation, etc.
- Social risks
  - ➢ Biased treatment, spread of misinformation, etc.

**Direction of AI Safety Evaluation**

- Examining the acceptability of newly emerging risks in relation to the benefits of AI robots
- Designing evaluation methods that combine real-world validation with simulated scenarios

* A pilot evaluation is planned at KAWARUBA, a co-creation hub for social innovation.

**SWG Roadmap**

| Short-term initiatives (FY2025) | Medium-term initiatives (FY2026–2027) | Long-term initiatives (future vision) |
|---|---|---|
| ● Identify AI safety risks based on interactions by robot type<br>● Design multilayered risk evaluation scenarios including psychological impact and social acceptability<br>● Explore risk mitigation measures and conduct preliminary trial evaluations in simulated environments | ● Verify the effectiveness of AI implementation in real-world settings<br>● Systematize risk indicators<br>● Establish and refine integrated analytical methods, incorporating factors such as acceptability and reliability | ● Systematize evaluation models and design guidelines<br>● Establish a common framework that enables the application of robots across a wide range of use cases<br>● Develop mechanisms for continuous evaluation and improvement during the operational phase |

# Data Quality SWG

- The Data Quality SWG aims to establish methods for visualizing and managing data quality, recognizing data as a foundation for AI safety.
- The SWG will develop practical guides and evaluation tools to complement existing standards and apply them across sector-specific SWGs and other partners.

**Current Status and Issues**

- Defects in input data can lead to unintended outputs
  - ➢ Requires the establishment and implementation of methods to visualize and manage data quality
- Although multiple international standards and theoretical frameworks exist,
  - ➢ Practical evaluation methods and implementation procedures remain underdeveloped
  - ➢ Many sites face challenges in practical adoption
- Data Quality Management Guidebook Ver.1.0 (AISI, Mar. 2025)
- Practical and sufficient data quality management is essential

**Approach to Ensuring Data Quality**

- Development of practical guides
  - ➢ Organizing evaluation perspectives and practical frameworks
  - ➢ Updating the guidebook
- Provision of evaluation tools
  - ➢ Use as a first step in data quality management
  - ➢ Lowering barriers to practical implementation
- Implementation and feedback
  - ➢ Trial evaluation of data quality based on real-world business scenarios

**SWG Roadmap**

| Short-term initiatives (FY2025) | Medium-term initiatives (FY2026–2027) | Long-term initiatives (future vision) |
|---|---|---|
| ● Conduct an application verification based on the "Data Quality Management Guidebook (Version 1.0)"<br>● Develop and evaluate simplified prototypes of quality evaluation tools (e.g., checklists)<br>● Conduct pilot evaluations to test both their ease of adoption and practical effectiveness | ● Conduct implementation demonstrations across sectors to establish practical utilization models<br>● Support multimodal data formats and emerging methods such as multi-agent systems<br>● Expand functions of evaluation tools in full scale | ● Establish operational models for continuous data quality evaluation in organizations<br>● Develop mechanisms for updating the guides and evaluation tools in response to technological advancements and societal changes<br>● Reflect Japan-originated data quality evaluation frameworks and metric definitions to global discussions |

# Conformity Assessment SWG

AISI Japan AI Safety Institute

- The Conformity Assessment SWG aims to establish AI conformity assessment methods, including self-declaration of conformity, through the activities of sector-specific SWGs.
- In collaboration with the Data Quality SWG, the SWG will explore composite evaluation approaches that consider the consistency of AI outputs and their context of use.

**Current Status and Issues**

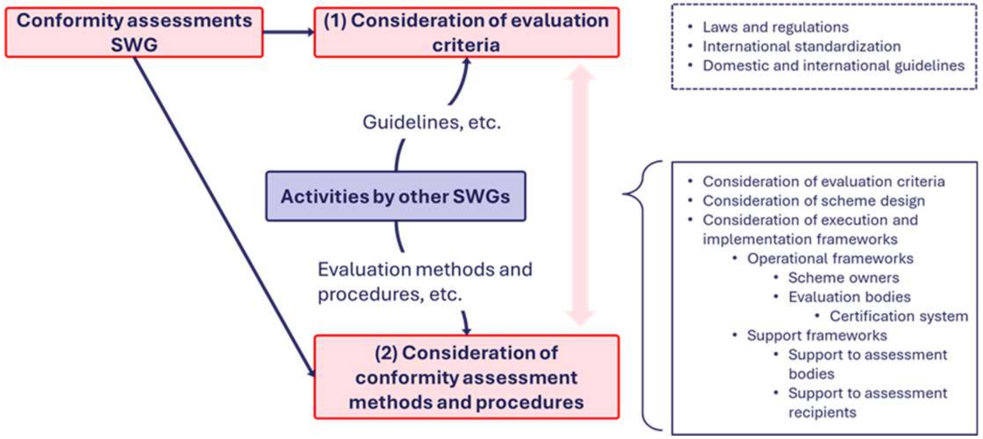
- Traditional conformity assessment frameworks cannot keep up with rapidly evolving AI technologies.
- ISO/IEC 42007, now under development, offers a high-level framework for conformity assessment for AI.
  - ➢ Flexible, appropriate evaluation methods and practical procedures are needed.

**Approach to establishing conformity assessment methods**

- Extract AI system requirements based on the activities of sector-specific SWGs

**Steps**: Hearing / Issue Identification | Scheme Design | Tool and framework development | Alignment with International Standards

- Explore composite evaluation methods in collaboration with the Data Quality SWG


Conformity assessments SWG
(1) Consideration of evaluation criteria
Guidelines, etc.
Activities by other SWGs
Evaluation methods and procedures, etc.
(2) Consideration of conformity assessment methods and procedures

- Laws and regulations
- International standardization
- Domestic and international guidelines

- Consideration of evaluation criteria
- Consideration of scheme design
- Consideration of execution and implementation frameworks
  - Operational frameworks
    - Scheme owners
    - Evaluation bodies
      - Certification system
  - Support frameworks
    - Support to assessment bodies
    - Support to assessment recipients

| SWG Roadmap | Short-term initiatives (FY2025) | Medium-term initiatives (FY2026–2027) | Long-term initiatives (future vision) |
|---|---|---|---|
| | ● Collaborate with sector-specific SWGs to conduct interviews about evaluation needs and the current conditions of assessment recipients<br>● Form a new SWG composed primarily of relevant organizations involved in conformity assessments, and analyze existing conformity assessment methods | ● Design conformity assessment schemes to ensure comprehensive assurance and accountability, based on the international standardization discussions<br>● Engage in the elaboration and trial implementation of a conformity assessment scheme for AI to verify its effectiveness | ● Develop guidelines aligned with international implementation models<br>● Work on the phased deployment, with the goal of establishing a practical, internationally interoperable framework |

AISI

Japan AI Safety Institute