Business Demonstration Project to Promote AI Utilization through AI Safety Evaluation

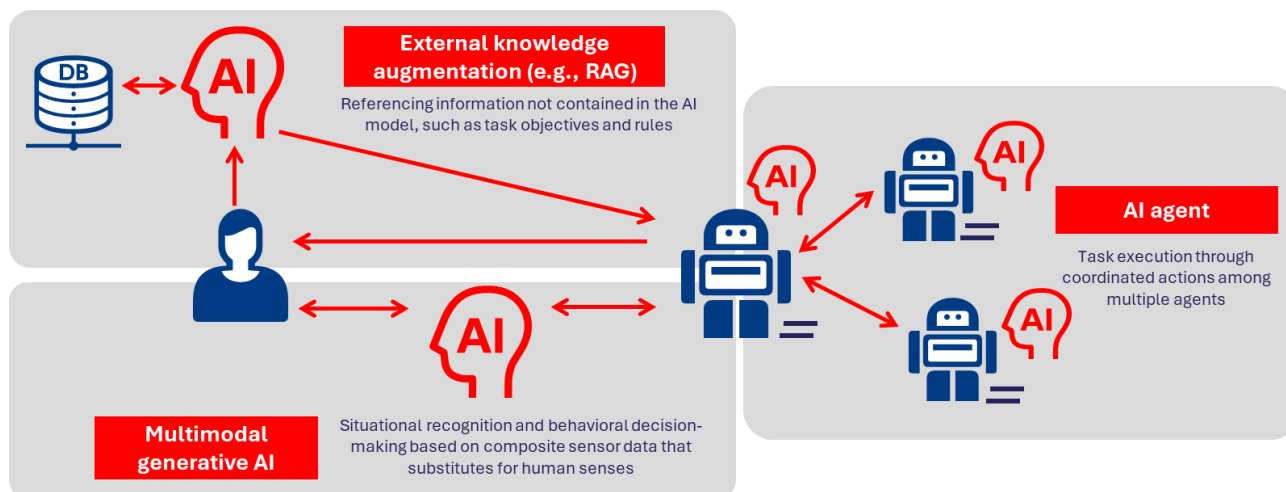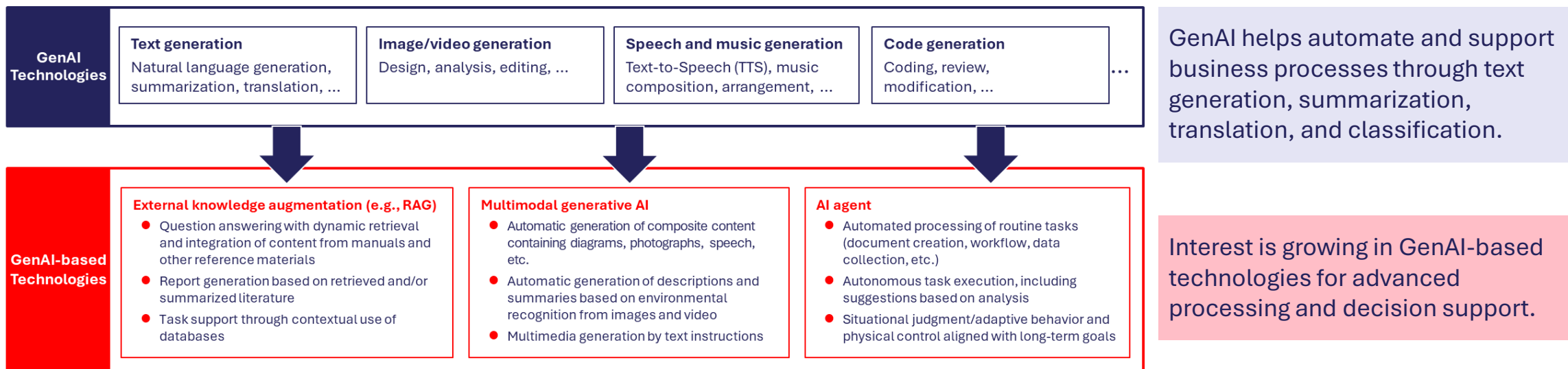**— Toward a Society Where AI Trustworthiness and Innovation Coexist—**

January 15, 2026

Japan AI Safety Institute

AISI Japan
AI Safety Institute

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	Image/video generation	Speech and music generation	Code generation	
Natural language generation, summarization, translation, …	Design, analysis, editing, …	Text-to-Speech (TTS), music composition, arrangement, …	Coding, review, modification, …	…

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

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

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

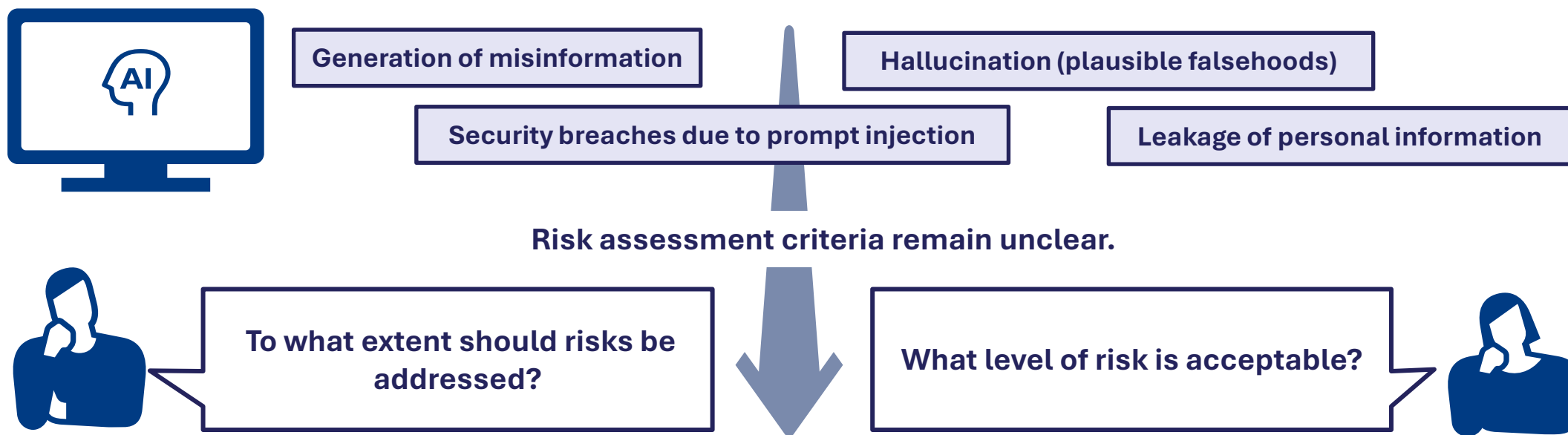Beyond digital tasks, AI is entering the physical world, supporting behavior control and operations.

Examples include:
- Drones for disaster search
- Robots for coordinated transport in factories

A new phase of AI as social infrastructure

2

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

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.".
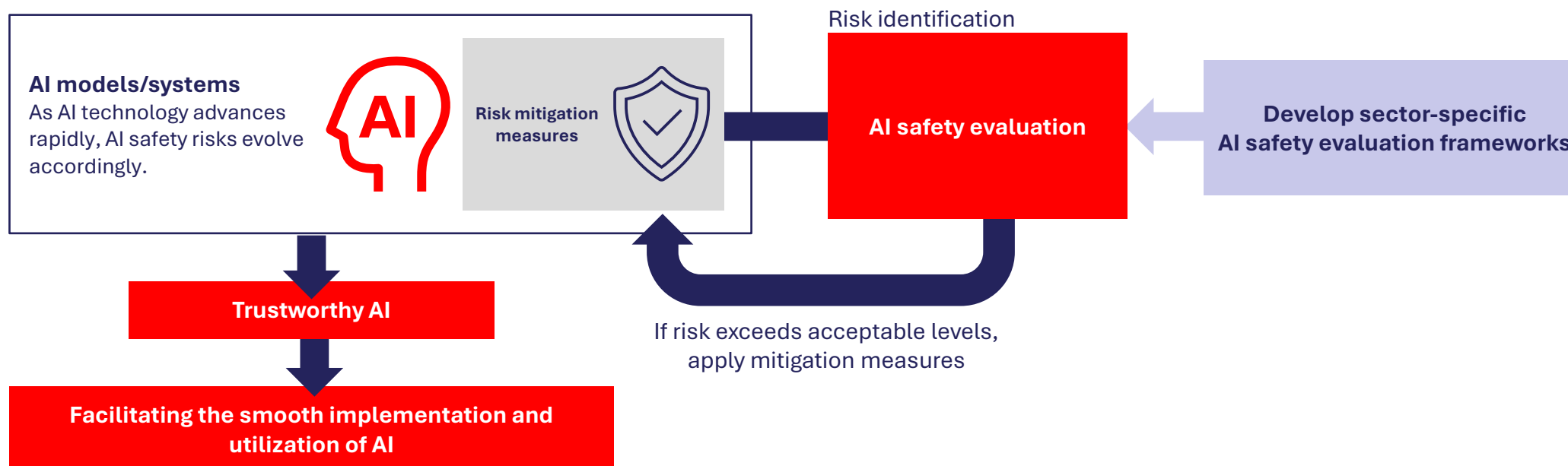
3

Objectives and Goals of the Business Demonstration WG

- 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.
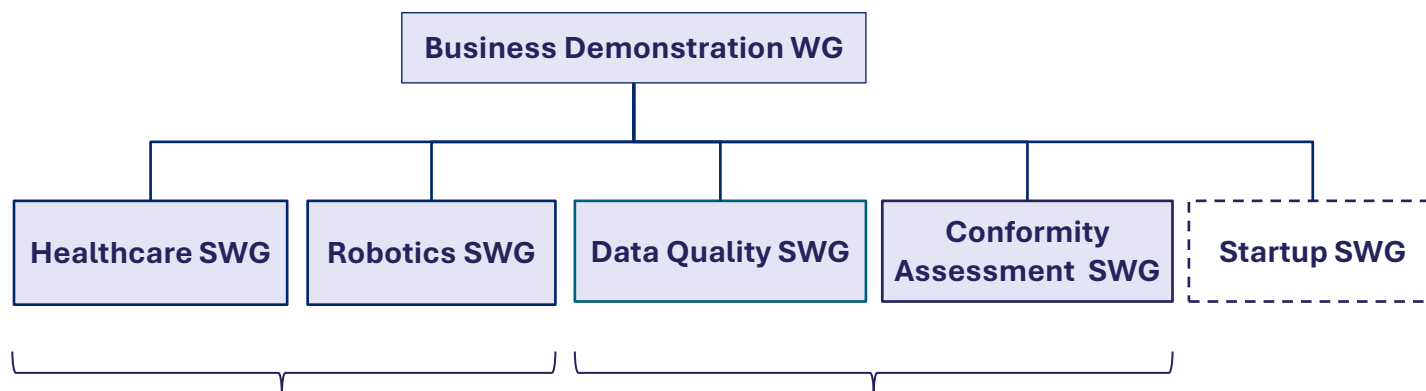
Medium- to long-term goals	Through the development of an AI safety evaluation framework: (1) to create opportunities that contribute to solving societal challenges—such as those related to healthcare, labor, and aging—by promoting the smooth introduction and widespread adoption of AI across industries, thereby supporting its societal implementation; and (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.

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

4

Structure of the Business Demonstration WG

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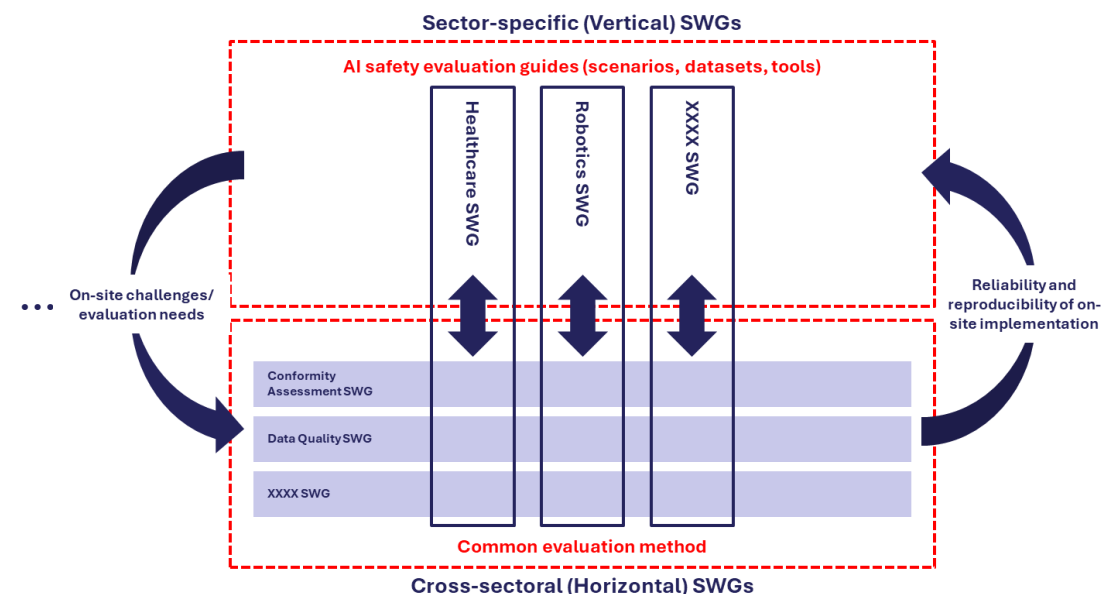- In FY2025, it consists of four SWGs: Healthcare, Robotics, Data Quality, and Conformity Assessment.

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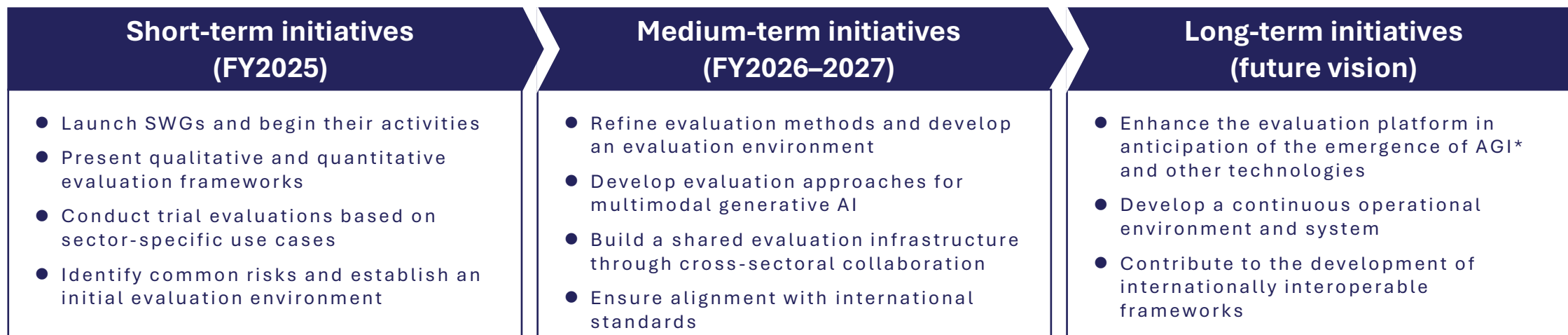
Cross-sectoral (Horizontal) SWG: Develop cross-sectoral evaluation methods

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

- 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.

5

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)	Medium-term initiatives (FY2026–2027)	Long-term initiatives (future vision)
• 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	• 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	• 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

AISI

Japan AI Safety Institute