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

AISI Japan AI Safety Institute

Japan AI Safety Institute
Business Demonstration Working Group

Table of Contents

0. Purpose of This Document

This Vision Paper aims to serve as both an agenda and a guideline for realizing a society in which AI 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, in coordination with the demonstration activities of the Japan AI Safety Institute (AISI).

The definition of "AI Safety" in this Vision Paper is based on the "Guide to Evaluation Perspectives on AI Safety," published by AISI in September 2024. In that guide, "AI Safety" is defined as a "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.)"

Appropriate evaluation of AI safety is an essential element in achieving coexistence of trustworthiness and innovation by demonstrating that AI is a trustworthy technology that forms a foundation for AI to be accepted in society as a "trustworthy AI.".

We hope that this Vision Paper will broadly serve as a useful reference for business operators, researchers, policymakers, and others who are working toward the safe and effective use of AI—for example, when designing frameworks and evaluation items related to AI safety. In particular, the Vision Paper is intended to be used by AI developers and providers (as defined in the "AI Guidelines for Business," established by the Ministry of Internal Affairs and Communications (MIC) and the Ministry of Economy, Trade and Industry (METI) in 2024) when evaluating the AI safety of foundation models, products, and services that are developed or used by the providers themselves. The Vision Paper is also expected to serve as a reference for user companies and municipalities that are introducing or utilizing AI, helping them identify and organize issues related to AI safety and understand the factors involved in its social acceptability. Additionally, the Vision Paper is expected to inform discussions among policymakers, certification bodies, and others toward the development of future evaluation guidelines. This Vision Paper aims to contribute to fostering a common understanding of AI safety and promoting its social implementation by serving as a reference for a wide range of stakeholders involved in evaluation, each from their respective perspectives.

1. Introduction

In recent years, the rapid advancement of generative AI and other AI technologies has brought significant changes to conventional business processes across a wide range of sectors, including industry, government, healthcare, and education. In particular, natural language processing technologies—exemplified by large language models (LLMs)—have significantly expanded their applicability to human decision-making and creative tasks, drawing unprecedented attention for their potential to generate both societal and economic impact. Furthermore, AI models capable of handling not only text but also images, speech, and other modalities in an integrated manner can serve as a common platform for a wide range of AI applications. Looking ahead, much of society's digital infrastructure and services may come to rely on such models, upon which individual AI services and agents could be built. In Japan, where labor shortages stemming from a declining birthrate and aging population have become an urgent issue, AI is seen as a key solution. Promoting its utilization is considered essential for the sustainable growth of both society and industry.

At the same time, as AI increasingly influences human behavior, judgment, and even public decision-making, new risks have emerged—such as the generation of misinformation and biases, the black-box nature of AI systems, and the difficulty of providing explanations. Since these risks may have a significant impact on businesses, organizations have taken a cautious stance toward AI adoption, As such, AI utilization has not yet progressed sufficiently. In a situation where convenience and efficiency are advancing alongside growing uncertainty and societal risks, ensuring AI safety while promoting its use is a shared and pressing challenge for the international community, including Japan. Addressing this challenge requires the implementation of appropriate risk management in parallel with the promotion of AI-driven innovation.

In response to this situation, the Japanese government established the Japan AI Safety Institute (AISI) in February 2024 as an organization responsible for reviewing and promoting evaluation methods and stans for "AI safety," with the aim of appropriately addressing the risks of AI while accelerating innovation. In September of the same year, AISI formulated and published a series of guidelines outlining its fundamental concepts. Furthermore, in March 2025, AISI established the "AI Safety Evaluation Working Group (Business Demonstration WG)" to promote and operationalize the guidelines. Building on the international agreement at the G7 Hiroshima AI Process, the "AI Guidelines for Business" (issued by the Ministry of Internal Affairs and Communications (MIC) and the Ministry of Economy, Trade and Industry (METI)), and trends in international AI standards (e.g., ISO/IEC JTC1 SC42), AISI and the WG have begun systematizing practical, use-case-based evaluation guidelines—including evaluation scenarios, datasets, and tools—through demonstration activities. The Business Demonstration WG brings together a wide range of stakeholders, primarily from the private sector, to identify use cases and needs from practical perspectives, share business insights, discuss challenges, and organize the findings into a systematic framework. By leveraging the agility of private businesses, the WG is expected to advance the development of practical rules through the formulation of evaluation guidelines based on the perspectives of industry and users.

The activities of the Business Demonstration WG aim to establish AI safety evaluation practices that contribute to the social implementation and industrial deployment of AI, with the goal of supporting not only government agencies but also private-sector businesses and engineers. They also aim to establish a Japan-originated, practical, and multi-layered evaluation framework by developing cross-sectoral evaluation methods and mechanisms for ensuring data quality and conformity assessment, while taking into account the differing risk structures and evaluation needs across sectors. These efforts are also envisioned to contribute to the creation of international and standardized rules.

2. Vision of the Business Demonstration WG

To accelerate the social implementation of AI, it is essential to conduct verification through both risk assessment and demonstration based on actual business use cases. The Business Demonstration WG is envisioned as a forum for supporting such activities. In particular, as the use of generative AI for tasks involving advanced judgment and output progresses, there is an increasing need to establish evaluation methods through practical use cases specific to each sector and function.

Against this backdrop, this chapter outlines the direction of the Business Demonstration WG's activities toward building an AI safety evaluation framework that supports the use of "trustworthy AI."

2.1. Current Landscape of AI Utilization

In recent years, AI technology has advanced rapidly. In particular, the emergence of large language models (LLMs) and generative AI is fundamentally transforming conventional approaches to AI utilization. ChatGPT, which was released at the end of 2022, has been used for a wide range of purposes—from everyday conversation to business support and creative activities—helping to create an environment in which everyone, from the general public to industry, can benefit from AI. Generative AI is contributing to the automation and support of many business processes through natural language generation, automatic summarization, translation, and categorization. In recent years, applied technologies that leverage these fundamental natural language processing capabilities have been attracting attention for enabling advanced information processing and decision support in real-world business operations. (Figure 2-1)

One of the key applied technologies of generative AI is retrieval-augmented generation (RAG), a representative method of external knowledge augmentation. This method allows AI to retrieve necessary information from external databases in real time and generate output based on the results. RAG is increasingly being implemented, particularly in FAQ response systems and business support tools that leverage internal corporate knowledge, thereby contributing to the development of domain-specific AI. Another important applied technology is multimodal generative AI, which processes and generates multiple modalities—such as text, images, and speech—enabling more flexible understanding and expression of information. By combining different sources of information, multimodal generative AI can achieve more accurate contextual understanding and situational judgment, enabling cognitive processing more akin to that of humans. Real-world applications are expected, for example, in generating explanatory texts based on medical images and in robots that integrate data from various sensors to replicate human perception and decision-making. Additionally, there has been significant progress in AI agents that can autonomously perform tasks and make decisions and take actions based on the given context. AI agents can largely be categorized into two types: the "workflow type," which processes tasks according to predetermined objectives and workflows; and the "autonomous type," which makes decisions and takes actions independently in response to changes in the external environment. Workflow-type AI agents are already being widely used as automation tools for routine business processes, contributing to greater operational efficiency and standardization. On the other hand, autonomous AI agents are expected to be applied not only in digital spaces—for example, to support decision-making and provide automated responses in complex information environments—but also in physical spaces, where they can be used for behavior control and operational support. Examples include drones used for search operations at disaster sites and systems in which multiple robots collaborate to transport materials within factories. Implementation in various real-world settings is becoming increasingly realistic. (Figure 2-2)

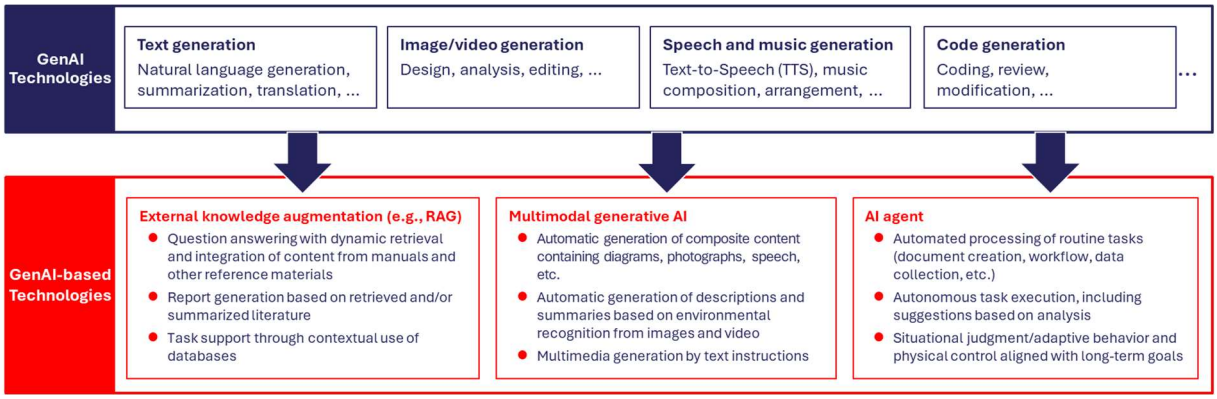

Figure 2-1: AI technology classification and application map

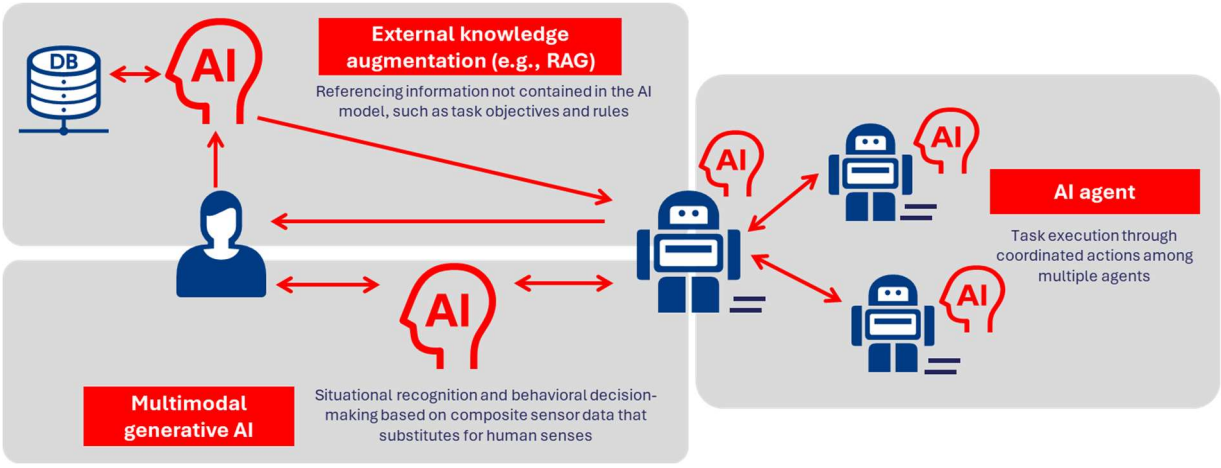

Figure 2-2: Utilization examples of generative AI-based technologies in physical spaces

As these cases illustrate, the advancement of generative AI-based technologies is bringing about qualitative changes in the societal role of AI and its positioning within systems.

Traditionally, AI was introduced as an auxiliary element responsible only for limited judgment or processing. Today, however, it is increasingly being implemented as a key component that handles entire processes—including information acquisition, integration, and interpretation—while collaborating with other systems and humans to make decisions. As a result, AI is evolving from a stand-alone module into an integrated entity that works in coordination with multiple functions to make judgments and take action. In practice, AI applications are rapidly expanding beyond documentation and conversational tasks into real-world domains such as healthcare (e.g., generation of discharge summaries and supporting patient explanation), manufacturing (e.g., preventive maintenance using patrol inspection robots integrated with anomaly detection AI), and retail and customer service (e.g., service robots that move through facilities while providing guidance and engaging in conversation).

In addition, the potential of AI to provide complementary support is growing, even in areas traditionally dependent on human intuition and experience. Technologies capable of handling elements that were once considered difficult to mechanize—such as ambiguous judgments, sensory understanding, and flexible responses—have emerged, and their use at the field level is becoming a reality. These developments indicate that AI is entering a new phase in which it functions as social infrastructure.

In response to such technological progress and the changing role of AI, rulemaking efforts—such as the “AI Guidelines for Business”—are progressing in the industrial sectors and the international communities. At the same time, discussions on identifying risks and clarifying responsibilities are advancing in earnest, with the aim of promoting innovation while ensuring the use of “trustworthy AI.” This is laying the groundwork for developing concrete approaches to AI safety evaluation.

2.2. Necessity of AI Safety Evaluation

As the utilization of AI rapidly expands and brings new benefits, numerous risk events that could undermine public trust are also emerging. More specifically, there have been actual incidents involving the generation of misinformation, security breaches through prompt injection, and leakage of personal information. There are also reports of cases in which outputs generated by generative AI have damaged corporate reputations, as well as overseas cases in which the use of generative AI-based chatbots has had negative effects on human health or even life.

Furthermore, the output of generative AI often contains so-called "plausible falsehoods" (hallucinations). These arise from a combination of factors, including the nature of the model, its training methods, and the quality of the training data. As such, it is difficult to prevent them entirely.

To address such risks, it is essential to properly understand their nature and impact through AI safety evaluations. Based on this understanding, a realistic approach is to apply functional risk management—primarily through technical measures—such as leveraging external information retrieval (e.g., RAG) and clearly citing sources that underpin the output generated by AI.

On the other hand, the lack of clear benchmarks for determining the necessary level of action for safe AI use has led many organizations to hesitate in adopting AI. While efforts to identify and address risks are progressing, the absence of a framework for evaluating these efforts and determining the level at which they are considered socially acceptable remains a major barrier to implementation.

For AI to gain broad social acceptance, these risks need to be managed to a socially acceptable level, rather than attempting to eliminate them entirely. Therefore, there is an urgent need to develop a framework for AI safety evaluation that reflects both the application and risk characteristics of AI.

2.3. Policy Responses and Sector-Specific Initiatives

2.3.1. Related Policy and Regulatory Trends

Currently, the development of policies and institutional frameworks related to AI safety is accelerating worldwide. While the policy orientations and regulatory strictness vary by country and region, there is a common tendency to emphasize "risk-based assessment," "ensuring transparency," and "clarification of responsibilities." (Table 2-1)

In Japan, the "AI Guidelines for Business" were published in 2024. This set forth a policy stance that emphasizes a cooperative governance framework primarily led by private-sector entities, along with strong support for implementation. In response, AISI released the "Guide to Evaluation Perspectives on AI Safety," which is based on the "AI Guidelines for Business," as well as the "Guide to Red Teaming Methodology on AI Safety," one of the evaluation methods. The supplement to the "Guide to Red Teaming Methodology on AI Safety" presents vulnerabilities identified in chatbot services that use enterprise RAG for internal business data utilization, along with measures to address them.

Furthermore, the AI Promotion Act (Act on Promotion of Research and Development and Utilization of Artificial Intelligence-Related Technologies) was promulgated in June 2025, establishing the responsibilities and basic measures of the entire country, including businesses, for promoting AI research development and utilization. In response to these developments, AISI will work through sector-specific sub-working groups (SWGs) to identify risks associated with individual use cases and to develop concrete methods for AI safety evaluation. It should be noted that many of the areas covered by the sector-specific SWGs—such as healthcare and robotics—are already governed by relevant laws and industry guidelines. Therefore, appropriate consideration must be given to ensuring consistency and complementarity with these existing legal/regulatory/institutional frameworks when designing AI safety evaluations.

Meanwhile, on the international front, the EU has taken the lead in enacting the AI Act, institutionalizing a risk-based tiered system of obligations—for example, requiring third-party conformity assessments (certification) for high-risk AI systems. In contrast, the United States has

taken a non-regulatory approach that emphasizes voluntary and documented practices involving AI, rather than legislation. Flexible, practical implementation is being promoted through public-private collaboration under the AI Risk Management Framework (AI RMF) developed by the National Institute of Standards and Technology (NIST).

In addition, international standards such as ISO/IEC 42001 for AI management systems and ISO/IEC 42005 for impact assessments have already been published. The development of new international standards is also underway, including ISO/IEC 42007—a high-level framework for conformity assessments—in which Japan is actively participating.

Designing AI safety evaluations in Japan requires flexibility aligned with the implementation process as well as rigor tailored to the nature and characteristics of risk. It is also essential to ensure that the evaluation methods are applicable across a wide range of industries and companies and are interoperable with international standards. These are key issues for the WG in future guideline development. A flexible and effective evaluation framework that reflects international regulatory trends will be key to Japan's AI safety strategy.

Table 2-1: Comparison of AI systems and assessment approaches in major countries

	Japan	EU	USA	UK	China
Regulatory methods/ Approaches	**Soft law Risk based**	**Hard law Risk based**	**Soft law Voluntary/ documentation**	**Soft law Principle/norm based**	**Hard law Management model based**
Principles	Principles of Human-Centric AI Society (2019)	Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems (2018)	The Blueprint for an AI Bill of Rights (2022)	National AI Strategy (2021) AI Whitepaper (A pro-innovation approach to AI regulation) (2023)	Ethical Norms for New Generation Artificial Intelligence (2021)
Guidelines	AI Guidelines for Business (2024)	Ethics Guidelines for Trustworthy AI (2019)	Artificial Intelligence Risk Management Framework (AI RMF) (2023)	Implementing the UK's AI Regulatory Principles Initial Guidance for Regulators (2024) Artificial Intelligence Playbook for the UK Government (2025)	White Paper on Trustworthy Artificial Intelligence (2021) Standardized Guidelines for Ethical Governance of AI 2023 (2023)
Policies, etc.	AI Promotion Act (2025)	AI Act (2024)	Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023) *Withdrawn in 2025	Artificial Intelligence (Regulation) Bill (Submitted to the Senate in 2025)	Interim Measures for the Management of Generative Artificial Intelligence Services (2023)

2.3.2. Initiatives by Sector

In addition to government-led rulemaking, voluntary efforts by private industry groups are also making progress in AI utilization. This "support for voluntary efforts by business operators" is one of the three main pillars of the "AI Guidelines for Business."

For example, in the healthcare sector, the Japan Digital Health Alliance (JaDHA) has developed the "Guide to the Utilization of Generative AI for Healthcare Providers"—a set of voluntary guidelines intended for service providers that use generative AI in the healthcare domain. The guide (Version 2.0, published in February 2025) serves as a practical resource for healthcare providers seeking to deliver services and products that incorporate generative AI. Additionally, the AI Mental Healthcare Association of Japan (AIMH) is taking the lead in developing guidelines—together with experts and business operators—on how to evaluate the use of generative AI in sensitive areas such as mental healthcare, from both ethical and technical perspectives. In April 2025, the "Draft Guidelines on the Utilization of Generative AI for Mental Healthcare Providers" was published, presenting principles that include privacy protection and consideration of psychological impact.

In the robotics sector, evaluation frameworks for human interaction with service robots have been organized and accumulated within the domains of Human-Agent Interaction (HAI) and Human-Robot Interaction (HRI), which explore users' psychological responses, behavioral characteristics, and the processes of forming social relationships. As AI continues to evolve, the scope of these evaluations is expanding. In addition, the emergence of end-to-end development models that integrate the entire process from environmental recognition to behavior generation is signaling fundamental changes in

the very way robots are designed.

Furthermore, in the financial sector, the Financial Data Utilizing Association (FDUA) formulated the "FDUA Generative AI Guidelines" (Version 1.0, published in October 2024), presenting a risk evaluation framework for the appropriate use of generative AI and practical considerations to ensure legal compliance and accountability. Additionally, financial institutions are strengthening their internal governance frameworks and developing internal rules, reflecting efforts to balance the use of generative AI with the need to ensure its trustworthiness.

In other sectors—such as education, manufacturing, and local government—guidelines are being developed to facilitate AI utilization in ways that reflect the unique characteristics of each sector, and evaluation perspectives are being clarified. Privately led efforts to ensure AI safety are steadily expanding across a wide range of industries and business types.

These privately led efforts can be positioned as efforts to fine-tune AI safety evaluations for specific sectors and contexts, in alignment with the "Guide to Evaluation Perspectives on AI Safety" and other resources developed by AISI, while maintaining a common baseline. Building on insights from these existing private-sector guidelines, the sector-specific SWGs, within the Business Demonstration WG, conduct repeated demonstrations and verifications, with the aim of refining and advancing practical evaluation criteria for each sector.

2.4. Project Purpose and Implementation Structure

2.4.1. Objectives and Goals

The underlying purpose of the Business Demonstration WG activities is to build a framework that supports the assurance of AI safety in the societal implementation of AI, through collaboration among industry, government, and experts. Specifically, the core goal of this WG is to develop a framework that goes beyond a simple checklist approach and enables multifaceted evaluation—such as explainability, conformity, and data quality—based on users' understanding of risks.

In addition, through the development of an AI safety evaluation framework, this WG has set the following as its medium- to long-term goals:

(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.

Such an evaluation framework is designed to clarify the model requirements for "trustworthy AI," and to enable AI developers and providers to ensure compliance with these requirements and to explain and demonstrate them transparently to users and regulators. The framework should also be structured in a gradual and flexible manner so that it encourages implementation in practice on-site and promotes quality improvement, without imposing excessive burdens or becoming a mere formality.

Building on the "Guide to Evaluation Perspectives on AI Safety" and other guides, the Business Demonstration WG is working to construct such a framework while identifying industry-specific evaluation items and coordinating with cross-sectoral areas (e.g., data quality and conformity assessments). The aim is to present an AI safety evaluation model that supports the societal implementation of AI.

This systematic framework goes beyond a "defensive safety" measure to mitigate risks, but aims to leverage Japan's strengths in sectors such as healthcare and robotics, as well as its competitive advantage in quality, cost, and delivery (QCD). In doing so, the Business Demonstration WG aims to take an international leadership role in developing evaluation methods that provide reasonable accountability for AI safety. By actively leveraging these frameworks as "offensive safety" measures to promote AI-driven innovation, the Business Demonstration WG aims to pursue a strategic approach

that enhances its international competitiveness. “Offensive safety” refers to a framework that accelerates the utilization and implementation of “trustworthy AI” by clarifying rational approaches and methods for keeping risks within acceptable levels and using them as means to fulfill social accountability. This approach helps to suppress costs associated with excessive safety measures and uncertainty. (Figure 2-3)

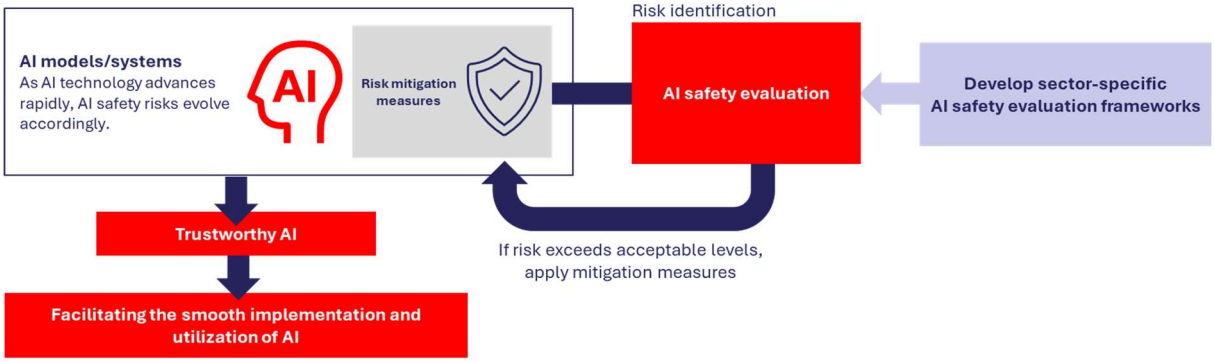

Figure 2-3: Overview of Business Demonstration WG activities

2.4.2. Implementation Structure of the Business Demonstration WG

The Business Demonstration WG has established multiple sub-working groups (SWGs) to develop effective AI safety evaluation frameworks aimed at realizing “trustworthy AI.” To be specific, the WG has established a structure that combines two types of SWGs: sector-specific (Vertical) SWGs, which work on implementation by developing evaluation guides—including evaluation scenarios, datasets, and tools—for each sector; and cross-sectoral (Horizontal) SWGs, which focus on developing a common evaluation method applicable across different sectors.

The sector-specific SWGs conduct business demonstrations related to AI safety evaluation in the healthcare and robotics sectors—areas where Japan has particular technological and industrial strengths and where ensuring AI safety is especially important as AI utilization advances—among sectors that are also of high international interest.

- Healthcare SWG:
  The healthcare sector has been at the forefront of AI utilization, with diverse services being developed—such as business support tools for medical institutions and chatbot services for general consumers. Some of these services have already demonstrated tangible contributions to solving specific issues in each domain. On the other hand, the healthcare sector involves the handling of highly sensitive information more often than other sectors. As such, clearly defining AI safety evaluation items and methods specific to the healthcare sector is expected to support the development of a safe and secure environment for AI utilization and to promote its societal implementation.
- Robotics SWG:
  Amid growing social demands stemming from labor shortages and the increasing complexity of on-site operations, the robotics sector has seen rising expectations for the implementation of AI. In particular, advances in vision-language-action (VLA) models, which integrate visual perception, natural language, and physical actions, are making it possible to automate physical tasks in a variety of environments based on natural language instructions. As these technologies are increasingly applied in scenarios involving human-robot interaction, ensuring safety in such contexts is becoming critically important. Therefore, further expansion in the use of AI can be expected by clarifying an AI safety evaluation framework that builds on the existing functional safety framework for robotics.
- Data Quality SWG:
  Today’s AI models are data-driven, learning from vast and diverse datasets. As a result, their performance and safety are heavily dependent on the quality of the data. Against this backdrop, it is essential to articulate evaluation perspectives and practical approaches to ensure data quality

as a foundation for AI quality.

- Conformity Assessment SWG:
  As international frameworks for AI risk management and accountability continue to take shape, there is growing interest in effective evaluation methods to confirm compliance. There is an expectation to propose a flexible and context-appropriate approach to conformity assessments that suits on-site needs.

With support from the WG Secretariat, each SWG organizes evaluation perspectives, develops guidelines and tools, and coordinates with other SWGs to ensure alignment, accumulating outputs as foundational resources for full-scale activities going forward. Additional SWGs will be established as needed.

The sector-specific (Vertical) SWGs and cross-sectoral (Horizontal) SWGs each play their respective roles, while complementing one another and working together to form an AI safety evaluation framework. In sector-specific SWGs, such as those for healthcare and robotics, specific on-site challenges and evaluation needs are identified based on actual use cases. These are then generalized in the cross-sectoral SWGs to help define evaluation items and guide the development of tools. On the other hand, the evaluation guidelines developed by the cross-sectoral SWGs are intended to be applied across different sectors and serve as guidelines to support on-site implementation. Through this bi-directional collaboration, the evaluation framework continues to evolve—balancing on-site specificity and broad applicability—and serves as a practical foundation for embedding AI safety evaluation in society. Furthermore, by presenting model cases and insights gained through demonstrations, the WG aims to contribute to international frameworks and related initiatives with Japan's original standards and methodologies. (Figure 2-4)

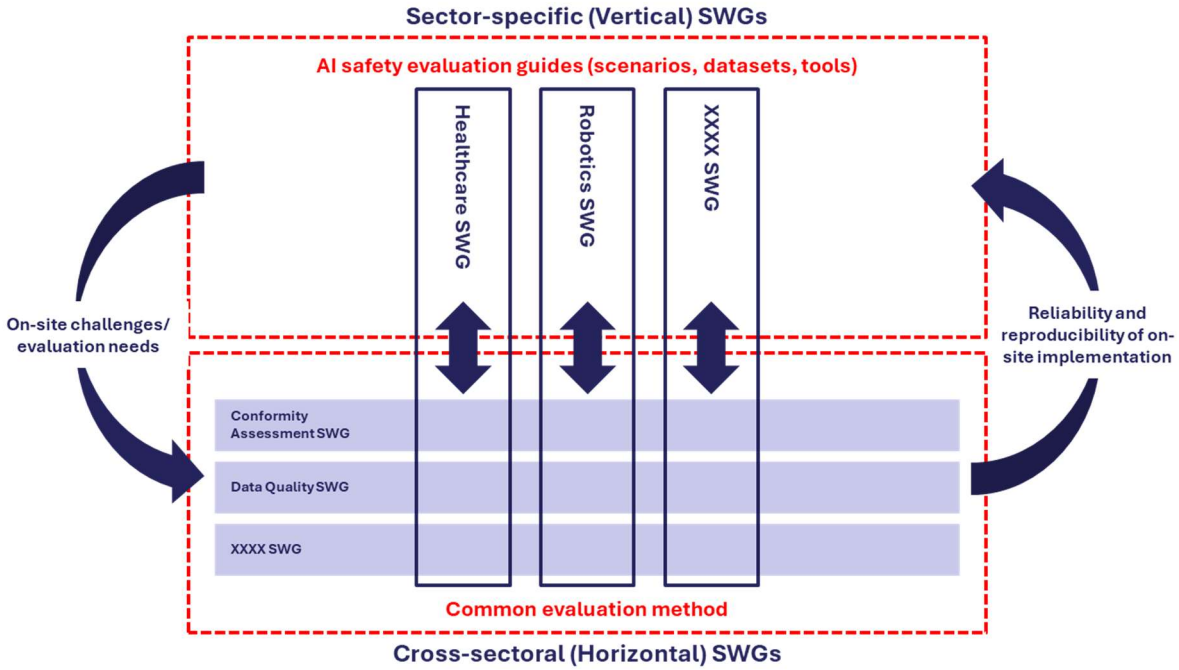

Figure 2-4: Overall vertical/horizontal structure of SWGs

2.4.3. Roadmap

The Business Demonstration WG will pursue the establishment of AI safety evaluation through a phased and continuous approach, with the aim of promoting the societal implementation of AI.

In FY2025, the WG will present evaluation frameworks that allow for both qualitative and quantitative assessment through the activities of each SWG. The SWGs will organize evaluation perspectives and conduct trial evaluations based on specific use cases in key areas such as healthcare and robotics. At the same time, they will identify risk elements common across sectors,

with the aim of establishing an initial evaluation environment for practical application in implementation settings.

In the medium term, based on evaluation tools, datasets, and other resources, the WG will work to establish a feedback loop through practical application and operation, refine evaluation methods, and develop an evaluation environment with shared access and verification functions. In particular, the WG will also address the increasing technological diversity of evaluation targets, including support for multimodal generative AI and emerging technologies accompanying the advancement of agent AI. The WG will also expand the scope of its SWG activities in phases by applying them to additional sectors with significant societal impact. While coordinating common issues and evaluation perspectives across sectors, the WG will promote ongoing collaboration and mutual use among stakeholders and further develop evaluation tools and data platforms as shared infrastructure. Through these efforts, the WG aims to enable application in diverse real-world settings while promoting alignment with international standards and fostering a shared understanding of evaluation methods both domestically and internationally.

In the long term, the WG aims to support the continuous operation of evaluation, while anticipating new AI safety risks that may arise with the emergence of artificial general intelligence (AGI) and other technologies. To that end, the WG will accumulate insights and enhance its platform, with view of coordinating with related regulatory systems. The WG will also continue to develop an operational environment that enables the integrated and continuous execution of evaluation, auditing, and operation, along with an evaluation system that can flexibly respond to discontinuous changes in technology and risk. These efforts will be continued with a view to full-scale adoption of AI across various sectors and domains. Furthermore, the WG will contribute to the development of internationally interoperable frameworks and operational rules by building on a Japan-originated evaluation framework. In doing so, the WG will aim to contribute to establishing global trust in AI safety evaluation and solidifying Japan's leadership in this field. (Figure 2-5)

Figure 2-5: Roadmap of the Business Demonstration WG

3. Visions of SWGs

This chapter outlines the efforts currently under consideration by each SWG (sector-specific or cross-sectoral) to develop a framework that ensures AI safety in the social implementation of AI. AI safety evaluations in business demonstrations require the evaluation of not only AI models but also AI systems. However, a comprehensive evaluation framework that includes this layer has yet to be established. To enhance the credibility of such evaluations and build a reliable framework for evaluating AI models and systems, it is essential to adopt an approach to consider AI safety for each sector and specific use case based on various AISI guides. In particular, the activities of the Business Demonstration WG should be defined so that they function in a mutually complementary manner when developing sector-specific AI safety evaluation guides, including evaluation scenarios, datasets, and tools. In addition, these activities should be based on advancements in each sector, such as healthcare, robotics, data quality, and conformity assessments. The objective is to provide a framework that establishes sector-specific consensus on AI safety conformity assessments, including self-declaration of conformity.

3.1. Healthcare SWG

As outlined in 2.4.2, the healthcare sector has been at the forefront of AI utilization. This includes the development of various services, such as business efficiency tools for medical institutions and chatbots for general consumers. Some of these solutions have already contributed to the solution of real-world challenges. However, compared to other sectors, the healthcare sector involves the handling of highly sensitive information. Therefore, it faces an urgent need to clearly define evaluation perspectives and methods for AI safety. This is expected to facilitate the establishment of a safe and secure environment for AI use and promote its social implementation.

This section provides an overview of the current state of AI utilization in the healthcare sector and outlines the key challenges and the roadmap of efforts for AI safety evaluation based on specific use cases.

3.1.1. Current Landscape of AI Utilization

In recent years, AI utilization has been rapidly spreading to the healthcare sector, driven in part by advances in natural language processing technologies enabled by generative AI.

Especially in Japan, the healthcare sector faces challenges such as an aging population and rising medical demand. These challenges, along with the implementation of the Work Style Reform laws in April 2024 and the resulting restrictions on overtime work at medical institutions, have accelerated digital transformation efforts within the sector. Medical institutions are increasingly adopting and utilizing digital technologies to improve operational efficiency and enhance productivity of healthcare professionals. In this context, generative AI has garnered attention as a promising solution to reduce administrative burdens (particularly documentation), allowing healthcare professionals to focus more on their core responsibilities such as patient care and communication. In Japanese healthcare settings, documentation tasks, such as preparation of discharge summaries, have been performed manually by healthcare professionals, contributing significantly to their workload. A survey found that "clerical duties, (including recordkeeping, report writing, and document organization)" are the second most common cause of doctors working overtime (59.8%), following "patient care."[1]

[1] Special site for Work Style Reform for Doctors
https://iryou-ishi-hatarakikata.mhlw.go.jp/about/

Additionally, the healthcare sector is experiencing growing demand for continuous, personalized self-care and self-medication, alongside the spread of products and services that can accumulate Personal Health Record data (e.g., wearable devices), and increasing public awareness of health and well-being. These trends are spurring the development of consumer products and services that leverage generative AI for personal health management (e.g., health management apps). As technologies targeted toward the general consumer become more affordable and convenient, their integration into everyday life is expected to rise. The AI healthcare market is projected to grow substantially, from USD 39.25 billion in 2025 to USD 504.17 billion by 2032.[2]

3.1.2. Importance of AI Safety Evaluation and Expectations for Promoting its Use

As noted above, the adoption of generative AI products and services is advancing in the healthcare sector, both within medical institutions and among general consumers. However, it is critical to evaluate not only their accuracy and convenience, but also their safe utilization. Specifically, there is a pressing need to establish an evaluation framework capable of predicting, controlling, and verifying the potential impact of information proposed or output by generative AI on medical institutions and general consumers.

Taking into account these needs and circumstances, possible use cases may include (Figure 3-1):

- [For medical institutions] Discharge summary preparation support: summarize medical records and generate document templates
- [For medical institutions] Patient explanation support: organize and present key points of diseases and treatment plans in natural language
- [For general consumers] Provision of health information: deliver guidance on diet and exercise based on Personal Health Records

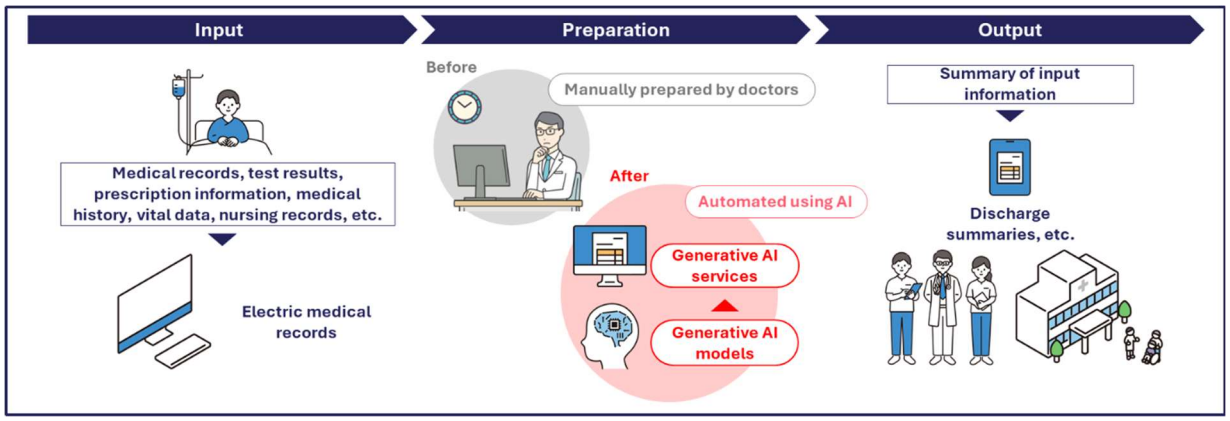

Figure 3-1: Examples of generative AI utilization in healthcare settings
(discharge summary preparation, etc.)

However, the healthcare sector inherently involves a high potential for directly affecting individuals' lives and physical well-being. Accordingly, concerns persist regarding issues such as hallucinations in generative AI outputs and the handling of sensitive information. As a result, the perceived "trustworthiness" of generative AI products and services has become a critical factor in determining their adoption and implementation. In response, the Japan Digital Health Alliance (JaDHA) published the "Guide to the Utilization of Generative AI for Healthcare Providers (Version 2.0)" in February 2025. This guide offers key checkpoints for healthcare providers to self-assess whether their generative AI-powered products or services may pose undue harm or disadvantage to users. This guideline has brought a certain degree of clarity for what aspects should be addressed when providing generative AI-powered products and services. However, evaluation items and methods for determining whether

[2] Fortune Business Insights
https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-in-healthcare-market-100534

such considerations have been sufficiently addressed to be deemed “trustworthy” have yet to be established. The absence of such established evaluation frameworks continues to raise concerns about the adoption and utilization of generative AI, presenting a significant barrier to the social implementation of generative AI in the healthcare sector.

Therefore, there is an increasing expectation for building a safe and secure environment for generative AI utilization by systematizing and clarifying evaluation methods and verification cases, tailored to specific use cases, and reflecting safety evaluation items to the design of generative AI products and services.

Against this backdrop, the Healthcare SWG will advance efforts to establish flexible evaluation perspectives that support the social implementation of generative AI in the healthcare sector. (Figure 3-2)

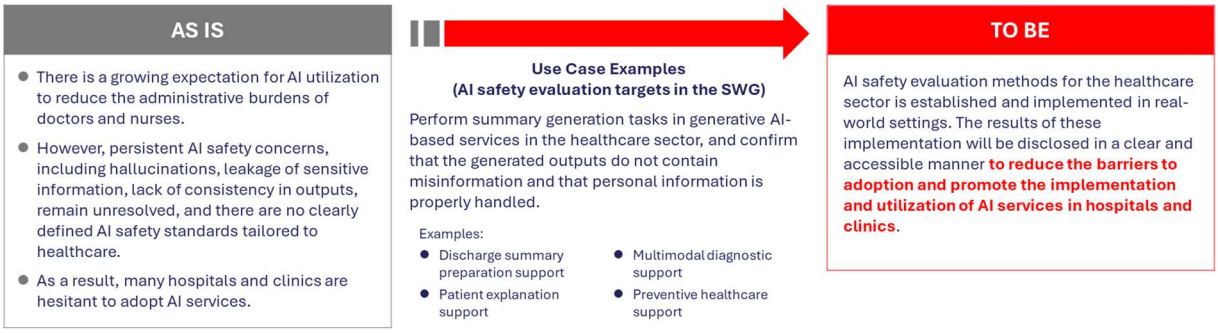

Figure 3-2: Overview of AS IS, Use Case Examples, and TO BE

3.1.3. Potential Risk Examples

The following outlines key risks and evaluation perspectives relevant to the healthcare sector. The Healthcare SWG intends to further organize and refine this list through discussions and considerations informed by use cases.

- Privacy leakage: Data containing personal information is not masked completely during input or output processes.
- Hallucination: Generation of texts containing factual inaccuracies may result in medical errors or the provision of incorrect information.
- Lack of output consistency: Variability in AI-generated outputs, even when provided with identical inputs, can undermine their reliability as records.
- Contamination bias: Biases in training data about age, gender, disease type, or others may lead to inappropriate outputs.
- Unclear accountability: Attribution of responsibility for outputs becomes unclear.
- Lack of explainability: An absence of transparent reasoning behind outputs can hinder understanding in medical settings.

To address these potential risks, the following perspectives should be included in AI safety evaluation:

- Is personal information handled appropriately?
- Are there safeguards to minimize hallucinations and prevent the omissions of critical information?
- Is there consistency in outputs?
- Is the network environment of medical institutions secure enough to support the safe implementation of generative AI services?

To promote the adoption of AI in the healthcare sector, it is essential to incorporate the necessary evaluation perspectives into each use case, develop both quantitative and qualitative evaluation models, and consider usability for healthcare professionals.

3.1.4. Direction of AI Safety Evaluation

The Healthcare SWG will advance the development of an AI safety evaluation framework based on practical needs and risk concerns, referencing existing guidelines such as the “Guide to Evaluation Perspectives on AI Safety” by AISI and the “Guide to the Utilization of Generative AI for Healthcare Providers” by JaDHA.

While generative AI outputs subject to the evaluation include text, images, speech, and multimodal formats, the initial focus will be on text-generating AI, given its current prevalence in the healthcare sector. The evaluation will target the AI safety of products and services not classified as medical devices or medical device programs (Non-SaMD: Non-Software as a Medical Device[3]). As generative AI technologies and services continue to evolve in the industry, the scope of these efforts may be updated as needed.

The following are examples of evaluation items. The SWG plans to further organize and refine these items through discussions and considerations based on specific use cases.

- Reproducibility of output: Confirm that a certain level of consistency is maintained in outputs generated for the same input.
- Detection of hallucination: Verify the presence and accuracy of mechanisms designed to prevent the automatic generation of false information.
- Privacy protection: Confirm that personal information is properly masked, especially in the input and output processes of generative AI.
- Output quality indication: Evaluate aspects such as style, readability, presence of typographical errors or omissions in summaries. Through both quantitative and qualitative evaluations, ensure the outputs meet quality standards appropriate for use in medical settings.

To enable practical implementation of these evaluation items, the SWG will pursue the development of both a checklist for AI safety evaluation and evaluation tools capable of simultaneously recording and comparing inputs and outputs. The SWG will also collaborate with medical institutions as well as AI developers and providers to apply these evaluation items and conduct feasibility assessments using actual data. The insights gained from these assessments are expected to enhance both the general applicability and on-site relevance of future evaluation frameworks.

3.1.5. Roadmap

This section presents the roadmap for the Healthcare SWG. Please note that the roadmap is subject to revision based on the progress of discussions and external developments, such as technical validations. Given that use cases in the healthcare sector are expected to include both B2B (e.g., for medical institutions) and B2C (for general consumers), the SWG will determine which use cases to address in the short term through discussions.

In FY2025, the SWG will identify representative use cases with a focus on Non-SaMDs and select those subject to consideration. The SWG will also conduct interviews with external experts to clarify the risk structure associated with generative AI utilization and design perspectives and scenarios necessary for AI safety evaluation. In addition, the SWG will work on the development of guides that organize evaluation perspectives that address key risks areas such as insufficient masking of personal information and hallucinations, , while also advancing consideration of evaluation datasets and tools.

In the medium term, the SWG will shift to the verification phase, where AI developers and providers will use the evaluation items, datasets, and tools developed in FY2025 to test their effectiveness. This phase will include pilot AI safety evaluations of generative AI products and services implemented in

[3] Non-SaMD refers to programs aimed at promoting health, etc., that are not subject to the Act on Securing Quality, Efficacy and Safety of Products including Pharmaceuticals and Medical Devices (PMD Act) in Japan.

multiple medical institutions and feasibility tests of the evaluation processes and tools outlined in the draft guides.

In the future, the SWG will engage in promotion and dissemination activities to encourage the use of the finalized guides, evaluation datasets, and evaluation tools within the healthcare sector. The SWG will also maintain and update the guides and other materials in response to the expansion of use cases. (Figure 3-3)

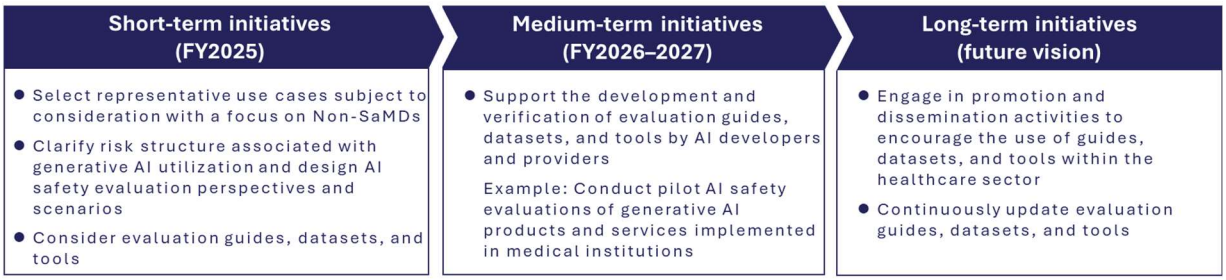

Figure 3-3: Roadmap of the Healthcare SWG

3.2. Robotics SWG

In the robotics sector, the behavior of AI-powered robots, ranging from service robots that operate in human environments to collaborative robots used in industrial settings, is becoming increasingly complex and flexible. With the growing prevalence of conversational robots and autonomous mobile robots, AI is playing a greater role across the full spectrum of functions: cognition, decision-making, and action. This expansion has introduced new safety concerns regarding the validity of AI-driven decisions and their impact on users. While traditional functional safety primarily addresses physical hazards, it is now essential to incorporate additional perspectives, including proactive risk mitigation, trust-building, contextual understanding in human-robot interaction, and support for humans. In this evolving context, the design and verification of AI safety evaluation, encompassing both conventional functional safety and the emerging dimensions of AI safety, has become an urgent issue in robotics.

This section provides an overview of the current state of AI utilization in the robotics sector and outlines the key challenges and the roadmap of efforts for AI safety evaluation based on specific use cases.

3.2.1. Current Landscape of AI Utilization

In recent years, the application of AI has been rapidly expanding in the robotics sector as well. AI has become particularly essential to the operation of both service robots and collaborative robots that require advanced decision-making and control across three fundamental dimensions: mobility, interaction, and task execution. These robots rely on AI not only for physical integration with sensors and actuators but also for performing the series of intellectual processes from “cognition” to “decision-making” to “action.” For example, mobile robots deployed in restaurants and airports are now expected to go beyond traditional control functions such as obstacle avoidance and path optimization. They need also to exhibit contextually appropriate behaviors, engaging in natural interactions with users, delivering appropriate instructions, and responding in a timely manner. In industry sectors, robots are increasingly being adopted in “long-tail” domains where the use of robots has traditionally faced limitations, such as food and pharmaceuticals, logistics, and small- to medium-sized enterprises. These use cases demand greater flexibility and ease of use, driving the adoption of AI-powered collaborative robots with greater versatility.

Amid this rising demand, the robotics AI market is projected to grow, with an estimated annual growth rate of over 26.8%, reaching USD 94.1 billion (approximately JPY 13.6 trillion) by 2031[4]. In particular, for robots with human-facing interfaces, such as life-support robots and communication robots, there is increasing interest in adopting generative AI and multimodal generative AI, including vision-language models (VLMs). These technologies are expected to allow robots to make decisions based on the five senses of humans, thereby realizing new types of user experiences. For collaborative robots, collaborations may extend from human-robot cooperation to robot-robot cooperation, where robots control and work with other robots. In recent years, the development of robot intelligence, called “Embodied AI,” has been advancing. It allows robots to integrate human dialogue with physical interaction, enabling deeper integration between situational understanding and action selection. (Table 3-1)

[4] Statista
https://www.statista.com/outlook/tmo/artificial-intelligence/ai-robotics/worldwide

Table 3-1: AI use case map in the robotics sector (conversational, mobile, and collaborative)

	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 – Provide product information to customers – Serve customers through natural conversation, etc. • Nursing and mental health care – Estimate emotions through voice and facial expression analysis – Show empathy and provide encouragement through dialogue, etc. • Tourist guidance – Provide tourist and route information to foreign visitors – Introduce recommended spots	• Autonomous delivery – Generate route dynamically – Identify environmental changes such as traffic congestion and change routes, etc. • Autonomous cleaning – Avoid obstacles – Detect dirt or wastes and determine appropriate cleaning methods, etc. • Security and patrol – Patrol facilities and check for abnormalities – Respond quickly when abnormalities are detected, etc.	• Understanding work instructions – Understand verbal instructions from humans – Convert given instructions into actions, etc. • Creation of procedures and manuals – Monitor tasks – Create procedures and manuals, etc. • Replication of human movements – Accurately recognize production techniques of skilled workers – Reproduce work styles, etc.

3.2.2. Importance of AI Safety Evaluation and Expectations for Promoting its Use

AI utilization in the robotics sector requires consideration of AI safety, taking into account communication and collaboration between humans and robots. In particular, for human-interactive robots, such as service robots and collaborative robots, it is necessary to evaluate not only functional safety which addresses traditional risks (physical hazards: collision avoidance, obstacle detection, etc.), but also the quality of human-robot interaction and context-appropriate decision making. Consequently, in the robotics sector, it is crucial to balance the benefits derived from the complex behavior of AI and robots with their inherent risks, and to develop a "safety evaluation" framework appropriate for social implementation. Currently, the following issues (AS IS) are being discussed:

- Conversational robots: Risks of giving overly authoritative responses or talking inappropriately to people with specific attributes.
- Mobile robots: Risks of misguidance or unintended entry due to route misrecognition; collision risks resulting from failures in environmental awareness.
- Collaborative robots: Risks of malfunctions resulting from loss of control leading to operational errors or serious accidents.

To address these issues, it is necessary to identify use cases based on real-world scenarios and clarify the necessary evaluation items from a safety perspective. In doing so, it is also necessary to consider whether it is use case intended for many and unspecified use by the general public (such as in public roads or public facilities), or for specific use within certain facilities.

For example, use cases include:

- Interaction and behavioral control support for the elderly and children
- Mobile robots recognizing high-risk environment such as crosswalks and performing autonomous path selection
- Multiple robots working collaboratively in restaurants, proactively avoiding collisions with humans (sequencing and avoidance)

In the future (TO BE), the development and publication of evaluation guidelines and case studies of uses cases will facilitate AI developers and providers in making informed decisions about the social implementation of AI-powered robots. This will, in turn, encourage the integration of AI safety considerations throughout the design, development, and operation phases, thereby lowering barriers to adoption for users and promoting AI utilization. (Figure 3-4)

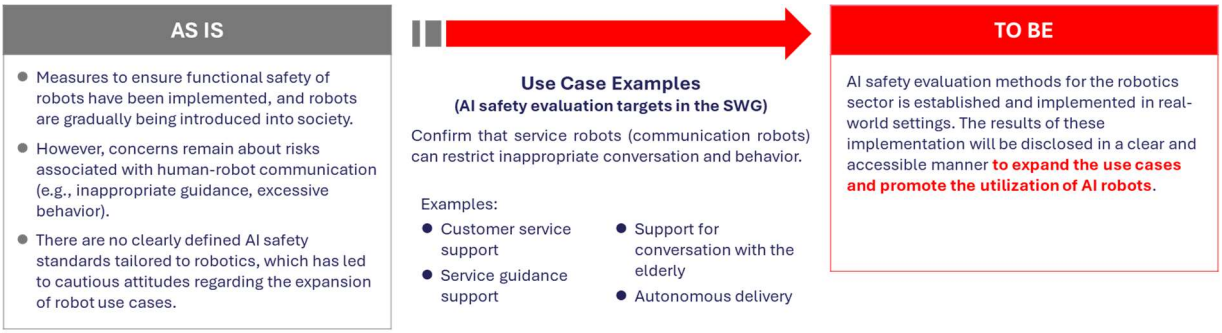

Figure 3-4: Overview of AS IS, Use Case Examples, and TO BE

3.2.3. Potential Risk Examples

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 that may arise from contextual human-robot interactions. (Table 3-2)

Table 3-2: Examples of AI safety risks in human-robot interactions

Category	Risk	Description	Examples
Physical risks	Collision, contact	Robot arms, wheels, or bodies collide or contact with humans	Malfunction of industrial robot arms, collision dues to route errors of mobile robots
	Entrapment, cuts	Fingers or hands get caught in moving parts	Collaborative robots losing control and coming too close to humans
	Electric shock, fire	Battery leakage, fire, conduction abnormality	Overheating during high-load AI training processes in service robots
Psychological risks	Fear, intimidation	Inappropriate behaviors	Loud noises, sudden movement, intimidating or nonsensical remarks
	Misunderstanding, overreliance	Stress or damage caused by task errors of robots	Poor judgment or improper work affecting operations
Social risks	Privacy violation	Collection and analysis of conversations, images, and voices	Recording conversation without consent; tracking of location information
	Bias, promotion of discrimination	Generative AI trained on skewed data producing biased responses	Inappropriate outputs related to gender, race, or age
	Inaccessibility, lack of fairness	Responses inappropriate for user attributes	Explanations too difficult for users to understand
	Loss of trust, spread of misinformation	Robots generating and providing incorrect information	Generation or dissemination of misinformation or fake news

The following are risks currently considered particularly important:

- Malfunction and misguidance: Navigation failures in narrow or multilevel environments, and inappropriate guidance based on misrecognition.
- Emotional and psychological impact: Inducing anxiety or distrust in users through intimidating or nonsensical remarks.
- Lack of accessibility and fairness: Unequal service delivery due to inadequate responses to users with diverse cognitive abilities or physical characteristics.

To address these risks, the SWG is currently exploring several key issues, such as clarifying functions and systems that are not appropriate for AI integration, defining safe development and evaluation processes, and developing evaluation frameworks specifically for interactions that integrate generative AI and multimodal AI (e.g., contextual appropriateness of behavior, systematization of speech risk indicators). The SWG aims to explore and identify emerging risks that may arise from the use of AI, explore acceptable levels of risks in line with the progress of social implementation and acceptance of AI, and establish new evaluation criteria that can articulate methods to control these risks. These efforts will be aligned with ongoing discussions and initiatives regarding traditional functional safety in the use of AI (e.g., international standardization efforts under ISO/IEC) which address high-risk scenarios such as physical harm to humans. Examples of indicators that are expected to become more important in future evaluation frameworks include the robot's ability to suppress actions that conflict with the user's intent and the consistency of robot responses,

regardless of user attributes.

3.2.4. Direction of AI Safety Evaluation

The Robotics SWG centers its efforts on the question: “What kind of risks, and to what extent, can humans accept from the use of AI, given the benefits that AI-powered robots offer?” To answer this question, the SWG is designing evaluations that combine evaluations in simulated environments (e.g., co-creation hub for social innovation KAWARUBA operated by Kawasaki Heavy Industries[5]) with scenario-based simulations. For example, the SWG aims to clarify safety indicators directly linked to behavioral design, such as speech control to prevent generative AI-powered robots making inappropriate remarks, and behavioral adaptations for specific user groups (e.g., speed limitations for the elderly, gaze control).

The technological components subject to evaluation are broadly divided into the following three categories:

- Sensing and cognition
  Examples: Accuracy of human recognition using vision sensors, noise resistance in speech recognition, integration and consistency of multimodal inputs
- Mobility and motion control
  Examples: Path stability in mobility algorithms, obstacle avoidance accuracy, responsiveness to environmental changes (e.g., sudden human movement), accuracy of motion prediction based on intent estimation
- Interface and explainability
  Examples: User interface that enables remote operators to accurately understand robot intent, presentation of explainable reasons behind decisions, development of decision support functions for users

Based on these technical components, risks will be classified as a matrix of “robot types (conversational, mobile, collaborative) x AI functions (cognition, decision-making, action, human operation support).” Moreover, AI safety standards can be considered from several perspectives, such as the use of AI in the development, evaluation, and operational processes, as well as hierarchical systems that include multiple robots and infrastructure. By developing AI safety standards appropriate for each case, the foundation for an evaluation framework that can accommodate future expansion of AI applications can be established.

When designing evaluation methods, it is essential to integrate both quantitative and qualitative approaches, with a focus on ensuring their mutual complementarity. For example, quantitative evaluations can be conducted by quantifying usage indicators, such as the frequency of inappropriate remarks and variability in response times. On the other hand, qualitative evaluations can be conducted by observing and analyzing users’ subjective views and sensibilities, such as impressions of robot’s remarks and acceptability of dialogue. The goal is to connect the two approaches within a unified evaluation framework, translating them into comprehensive indicators such as overall sense of security.

In addition, demonstrations in simulated environments (e.g., conversational robot experiments conducted at KAWARUBA) play a role in identifying areas for improvement in output control and behavior design. These findings will be used for the optimization of future robot behavior models and the development of implementation standards for output control, contributing to risk mitigation by guiding the establishment of AI safety guardrails and fallback mechanisms. (Figure 3-5)

The AI safety evaluation efforts undertaken by the Robotics SWG represent a starting point for systematizing safety evaluations tailored to the structure and task characteristics of robots. The SWG aims to build an evaluation model that not only defines the scope of acceptable risks but also

[5] CO-CREAION PARK KAWARUBA
https://kawaruba.com/

enhances social acceptance.

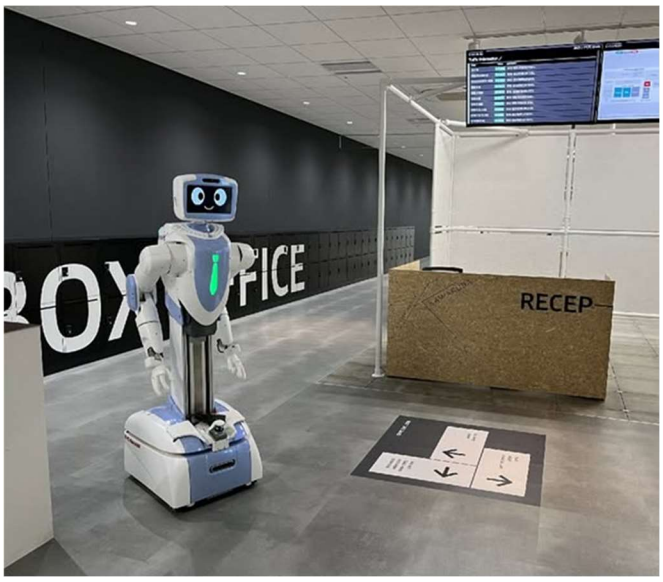

Figure 3-5: Demonstration environment and use case example at KAWARUBA

Source: Kawasaki Heavy Industries, Ltd.

3.2.5. Roadmap

The Robotics SWG will first focus on identifying AI safety risks based on interactions by robot type (e.g., conversational, mobile, collaborative) and then on designing multilayered evaluation scenarios. The SWG will also explore risk mitigation measures using simulated environments and conduct preliminary trial evaluations.

In the medium term, the SWG will verify the effectiveness of AI implementation in real-world settings. During the process, the SWG will also work on the systematization of risk indicators and the establishment and refinement of integrated analytical methods, incorporating factors such as acceptability and reliability.

In the future, the SWG will work on the further systematization and sharing of evaluation models and design guidelines to support broader social implementation. A key goal is to establish a common framework that enables the application of robots across a wide range of use cases. The SWG also aims to develop mechanisms for continuous evaluation and improvement during the operational phase and build a framework that can ensure sustained reliability in real-world settings. (Figure 3-6)

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

Figure 3-6: Roadmap of the Robotics SWG

3.3. Data Quality SWG

Ensuring the trustworthy AI systems fundamentally depends on the quality of data used across the training, evaluation, and utilization phases. Since AI has the “Garbage In, Garbage Out” nature, especially for generative AI that produces non-deterministic outputs, any defects in input data can lead to unintended outputs. Therefore, establishing and practicing effective methods to visualize and manage data quality is a critical challenge. However, there are some other issues. For example, international standards for data quality are varied, and it is difficult to comply with all of them. In addition, most of existing standards were developed for enterprise systems and structured data, and do not adequately address the data quality requirements of modern AI systems. This poses another issue, as the scope and complexity of data subject to quality management is dramatically expanding to include unstructured data (e.g., text, images), dynamically referenced knowledge bases, and third-party data.

Based on the understanding that managing data quality is an integral component of AI safety, this section outlines strategies to support the application of data quality management in practice, aligned with both domestic and international standardization efforts. Key focus areas include organizing evaluation perspectives, developing checklists and scoring tools, collecting feedback and driving improvements through real-world application, and defining the future directions of these efforts.

3.3.1. Current Status and Issues

AI is fundamentally a data-driven technology, and its safety is directly tied to the quality of data used in training, evaluation, and application. Data quality is a prerequisite for AI safety, as highlighted in the “Guide to Evaluation Perspectives on AI Safety” as one of the ten evaluation perspectives.

Traditionally, data quality management has centered on training and evaluation data, and has been an area of primary concern for AI developers and providers. However, the rise of generative AI has shifted the focus from development and provision of AI to the use of AI. With the advancements of architectures such as RAG and AI agents, dynamic referencing of internal and external data has become increasingly common. These usages are expected to serve as a mechanism to complement the pre-trained knowledge of AI and prevent hallucinations. However, if the referenced data is inaccurate or outdated, it may instead lead to erroneous outputs. As a result, managing the quality and provenance of referenced data has become as important as managing the quality of training data.

Several international standards have been proposed to address data quality. Representative standards include ISO 8000 (general data quality management), ISO/IEC 25012 (data quality characteristics of software), and ISO/IEC 5259 series (data quality in AI systems). These standards provide theoretical frameworks for data quality. In Japan, AISI published the “Data Quality Management Guidebook (Version 1.0)” in March 2025, covering key topics from a practical perspective. Efforts to bridge the gap between international standards and real-world practices are now underway.

However, as shown below, several technical and structural challenges remain unresolved:

- Gap between standards and practice
  Although there are several standards on data quality, they largely focus on definitions and lack specific evaluation methods or implementation procedures for practical use. For example, methods for evaluating concepts such as data accuracy and bias as measurable indicator and the timing of such evaluations are yet to be established. As a result, many organizations report facing obstacles during implementation, such as uncertainty about which standard to reference and how best to apply them in their organizations. Given the many standards, organizations seek to avoid excessive costs or overly strict governance and are demanding sufficient and reasonably priced approaches to data quality management.
- Data quality management in the AI era
  As standardization takes time, existing standards often lag behind emerging ways of using data, such as generative AI and RAG. Current challenges include limitations in applying

traditional statistical methods to the quality evaluation of unstructured data (e.g., text, images), lack of established methods to evaluate the quality of externally referenced data in RAG, and lack of clarity around practical approaches to ensure readability for machines and humans. In the face of the expansion of data handled by AI, the evaluation of data quality continues to rely on subjective judgment.

- Limited interest in data quality evaluation
  Another challenge is the lack of awareness about the importance of data quality among organizations and their executives. As such, the priority of efforts regarding data quality remains low. Because data is invisible and often seen as an intermediate component of a final outcome, the focus of attention tends to be on the quality of deliverables, such as AI models and analytical results. However, for the realization of "trustworthy AI," it is essential to raise awareness that data is the foundation that underpins AI safety.

3.3.2. Approach to Ensuring Data Quality

As outlined in the "Data Quality Management Guidebook (Version 1.0)," data quality management is structured around the eight phases of data lifecycle (data planning, data acquisition, data preparation, data processing, AI system, evaluation of output, deliver the result, and decommissioning). Data quality is ensured by systematizing quality evaluation checkpoints at each of these phases (Figure 3-7). In other words, data quality management means clarifying when, who, what, and how data should be evaluated, and ultimately establishing a "chain of trust," including data circulation to third parties. This approach for data quality management also requires comprehensive data management and governance mechanisms that span the entire data lifecycle.

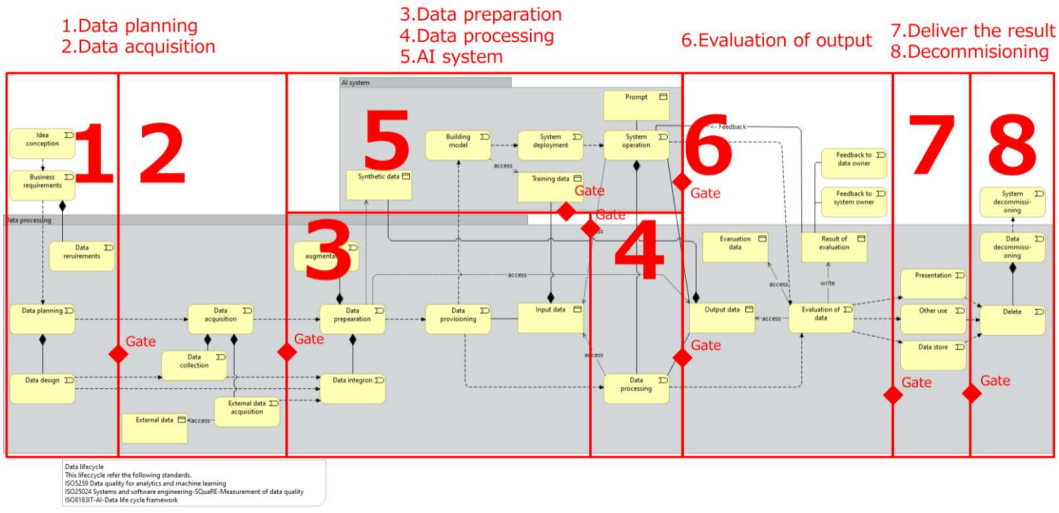

Figure 3-7: Data lifecycle including AI systems

The Data Quality SWG aims to lower the barriers to data quality management and promote its wide spread adoption across organizations with a practical approach. To achieve this goal, the SWG will develop practical guides and evaluation tools that complement existing standards. In addition, the SWG will present a comprehensive management framework that can address the data quality in different contexts, such as structured data quality in core systems, training data quality for machine learning, and reference data quality in the generative AI era. The goal is to facilitate the practical application of this framework.

To achieve this goal, the SWG will collect insights from data quality management experts (e.g., AI developers, providers, researchers) and pursue empirical activities through application verifications on-site.

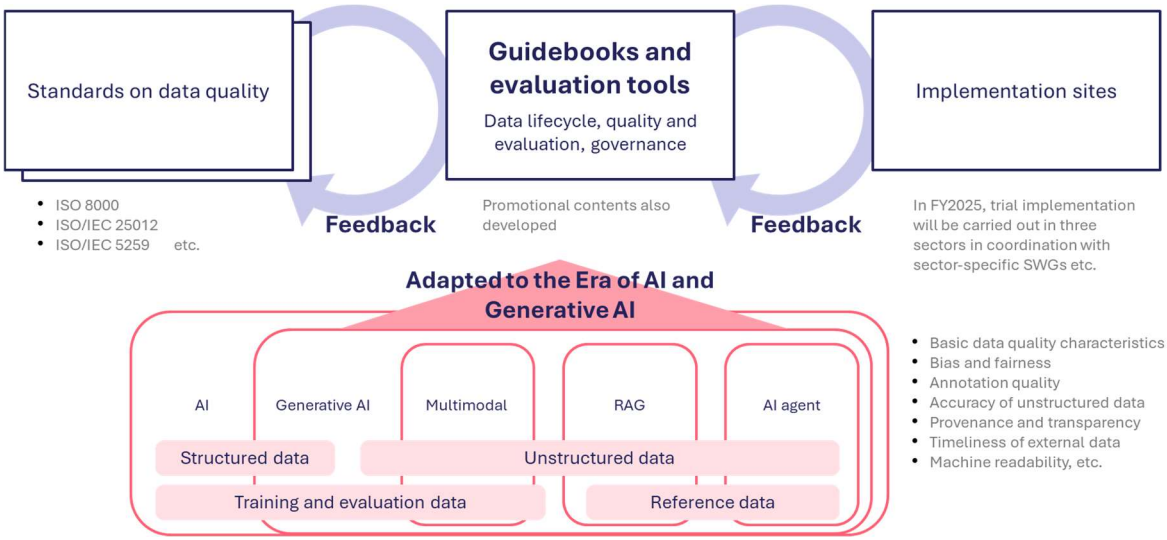

Figure 3-8: Data Quality SWG approach

Key efforts of the Data Quality SWG include:

- Guide development
  The SWG will update guidebooks that outline concrete evaluation perspectives and practical frameworks by phase and data type, while referring to and supplementing existing guidelines and standards. The SWG will also develop introductory versions of these materials and use cases that illustrate implementation procedures.
- Provision of evaluation tools
  To lower barriers to adoption, the SWG will provide evaluation tools that AI developers and providers can use as an entry point to data quality management. The following are use case examples of these tools currently under consideration:
  - Data subject to evaluation: Training data, safety evaluation data, reference data, etc.
  - Quantitative evaluation: Statistical distribution of data, missing data rate, bias metrics, etc.
  - Qualitative evaluation: Metadata availability, implementation status of quality management, etc.
  - Tool implementation: Checklists, simplified diagnostic tools, automated evaluation scripts (e.g., assumes lightweight implementation, such as Python applications). No special computing environments or advanced technical skills required.
  - Output format: Radar charts, score formats, or other visual formats that are intuitive and easy to use by on-site operators.

  In the future, the SWG will also explore advanced evaluation processing, automation, and sustainability.
- Application verification and feedback integration
  The SWG will conduct trial evaluations of data quality based on real-world operational scenarios in collaboration with sector-specific SWGs and other partners. Guides and tools will be updated based on evaluation results and on-site feedback. The SWG will also systematically gather and analyze application examples and issues occurred during the implementation process, support application in other sectors and organizations, and fill gaps between standards and operational practice in a phased manner.

Through these efforts, the SWG aims to establish a flexible quality evaluation framework capable of addressing diverse data formats and application scenarios. This framework will serve as a foundation for the social implementation of AI safety from a data quality perspective.

As a first step, the SWG will determine priorities for evaluation perspectives and output formats based on needs and feasibility, and continuously update the guides. In parallel, the SWG will use the

evaluation tools to concretize and reinforce the “data quality” perspective in the AI safety evaluation and apply them to verification scenarios, such as those of the sector-specific SWGs.

3.3.3. Roadmap

In FY2025, the SWG will conduct an application verification based on the “Data Quality Management Guidebook (Version 1.0),” collect and incorporate on-site feedback to refine implementation details and expand case studies. The SWG will also develop simplified prototypes of quality evaluation tools (e.g., checklists), and conduct pilot evaluations to test both their ease of adoption and practical effectiveness with sector-specific SWGs and other partners.

In the medium term, the SWG will conduct implementation demonstrations across sectors to establish utilization models across government and industry, and expand the scope of guides and evaluation tools to support multimodal data formats and emerging methods such as multi-agent systems. Additionally, as part of full-scale efforts to expand functions of evaluation tools, the SWG will advance the development of sector-specific templates and user interfaces for non-experts, explore frameworks for continuous data quality evaluation, and begin their empirical verification.

In the future, the SWG aims to establish an operational model for continuous data quality evaluation in organizations and establish it as a social infrastructure. The SWG will also develop mechanisms for updating the guides and evaluation tools in response to technological advancements and societal changes.

Furthermore, the SWG will reflect Japan-originated data quality evaluation frameworks and metric definitions to global discussions, in cooperation with international standardization bodies such as ISO/IEC JTC1 SC42. (Figure 3-9)

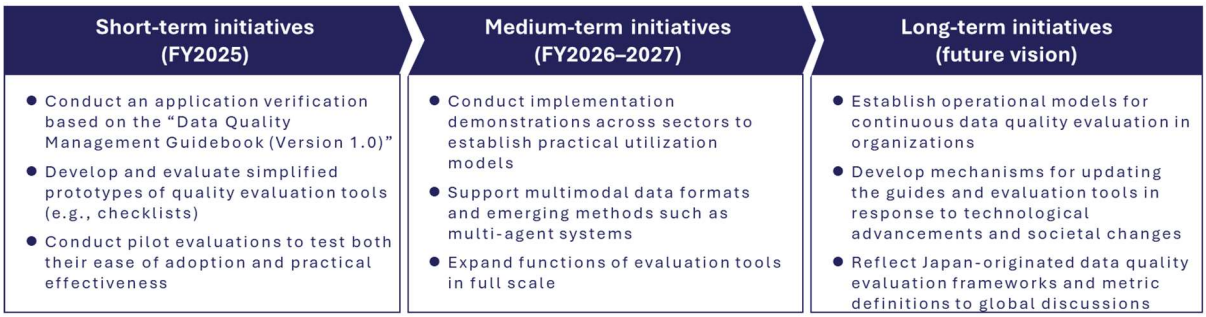

Figure 3-9: Roadmap of the Data Quality SWG

3.4. Conformity Assessment SWG

As AI technologies, service models, and operational frameworks continue to evolve dynamically, it is becoming increasingly difficult to apply traditional static and uniform conformity assessment frameworks.

ISO/IEC 42001, a new international standard for AI management systems, has been published as framework to guide AI utilization efforts of organizations. Similar to standards like ISO 9001 and ISO/IEC 27001, the new standard outlines requirements for organizational management systems and serves as a guideline for organizations to develop an AI management system. ISO/IEC 42001 is also expected to serve as the basis for third-party conformity assessment schemes, with preparatory work underway.

Development has also begun on ISO/IEC 42007, an international standard intended to provide a high-level framework for conformity assessments of AI systems. This emerging standard is closely related to the revision of ISO/IEC 17067, which defines the general framework for conformity assessment schemes. ISO/IEC 42007 is expected to support the design of flexible conformity assessment schemes for AI systems.

In this context, the SWG recognizes two critical challenges. Firstly, understanding international standards and effectively implementing them within organizations require interpretation and adaptation, and secondly, operating conformity assessments in accordance with international standards require elaboration of these standards into actionable processes. In light of these challenges, the SWG will conduct a feasibility study to clarify issues involved in developing concrete methods for conformity assessments for AI and establishing institutional and operational systems and processes. This section also discusses pilot evaluations conducted in collaboration with organizations involved in conformity assessments of AI systems (assessment bodies) and domestic AI developers and providers (assessment recipients), as well as a roadmap for future implementation.

3.4.1. Current Status and Issues

AI technologies are evolving rapidly. Given their dynamic and cross-cutting nature, it is difficult to evaluate modern AI systems with traditional static conformity assessment schemes, which are designed for vertical domains (products, people, organizations). In response to this challenge, ISO/IEC JTC1 SC42 is currently developing ISO/IEC 42007, which aims to define a high-level framework for conformity assessments of AI systems. However, further analysis is needed to understand the relationship between these new standards and the conformity assessment scheme based on the EU AI Act.

In Japan, effective adoption of ISO/IEC 42001 and ISO/IEC 42007 will require these standards to be elaborated at least at several stages. This includes the need to advance discussions on the development of practical conformity assessment methods that contribute to the growth of Japan's AI industry.

ISO/IEC 42007 serves as a high-level framework for conformity assessments of AI systems. Therefore, to develop concrete operational procedures and evaluation methods based on this international standard, its requirements need to be interpreted and elaborated into detailed, practical processes. This requires discussions with stakeholders. As for evaluation methods, flexible and appropriate methods and actionable procedures should be established through analysis and deliberation, including first-party conformity assessments (self-declaration of conformity) to third-party conformity assessments.

Assessment body's perspective

- Traditional conformity assessment frameworks often struggle to keep pace with rapidly evolving technologies such as AI. Therefore, it is necessary to develop evaluation methods that correspond to changing evaluation targets, reconsider the granularity of evaluations and validity criteria, and develop new conformity assessment methods to redesign the existing

frameworks.

- Traditional conformity assessment schemes, which focus on vertical domains (products, people, organizations), may not adequately or reliably address evolutionary technologies such as AI. This is particularly relevant for digital (system and service) domains, where the definitions and methods of conformity assessment remain under development. In addition, there is still technical immaturity in areas such as the development of evaluation methods, including the presentation (identification) of required evidence. Therefore, there is a need for developing flexible and possibly agile frameworks that can address AI safety (e.g., DevOps governance).
- To address the above issues, it is also necessary to develop conformity assessment methods including evidence design tailored to evaluation objectives and evidence-based conformance structure.

Assessment recipient's perspective

- Currently, there is no clear framework for companies to demonstrate that their AI systems conform to international standards or other global criteria. As a result, the burden and cost associated with undergoing conformity assessments remain unclear.
- There is an expectation for a practical conformity assessment scheme that allows for self-declaration of conformity, rather than relying exclusively on third-party assessments.
  - ➢ Assessment recipients are seeking a flexible scheme that enables them to demonstrate the conformity of AI systems they are developing or providing, whether through third-party assessments or self-declaration, without incurring excessive costs.

Perspectives for international collaboration

- Interoperability and international collaboration: To enable mutual recognition and interoperability on a global scale, it is essential to design flexible, phased schemes that are aligned with international standards.
- Collaborative structure: The successful design of conformity assessment schemes requires the establishment of a collaborative operational framework that involves a wide range of stakeholders, including evaluation bodies, AI developers and providers, and (if legislation is considered) policy-making authorities.

Taking these issues into account, the SWG will focus on developing conformity assessment methods that are robust enough for practical implementation and operation. These methods will form the basis for designing evaluation methods, procedural guides, and support tools tailored to the levels of AI systems.

3.4.2. Approach to Establishing Conformity Assessment Methods

The Conformity Assessment SWG, in collaboration with assessment bodies and assessment recipients, will work on the development of concrete conformity assessment methods, building on the works of ISO/IEC JTC1 SC42 and the efforts by AISI. The SWG aims to extract AI system requirements from specific efforts by the Healthcare SWG and Robotics SWG, which will utilize evaluation tools developed in line with AISI guides, and establish conformity assessment methods for AI (including self-declaration of conformity) and practical operational procedures.

To support the establishment, implementation, and eventual institutionalization of conformity assessment methods, the SWG will pursue the following concrete steps:

- Interviews and identification of issues: Identify issues with existing schemes and operations, and organize the needs of both assessment bodies and assessment recipients.
- Consideration for scheme design: Propose a draft of a flexible, evidence-based conformance structure tailored to specific use cases, and simulate phased implementation scenarios for

both third-party evaluations and self-declaration of conformity.

- Tools and system preparation: In coordination with the Data Quality SWG, explore the integrated application of output consistency and safety evaluation perspectives.
- Alignment with international standards: Reflect international developments, particularly those led by ISO/IEC JTC1 SC42, to the development and domestic implementation of evaluation guides.

The SWG will also explore the adoption of the Argumentation-based Approach as a framework for conformity assessments. This evidence-based, phased method validates conformity against specific evaluation goals, including safety. This method can be applied to targets that are continuously evolving, like AI.

The SWG, in collaboration with the Data Quality SWG, will also examine the consistency of AI outputs and composite evaluation methods, including context of use. (Figure 3-10)

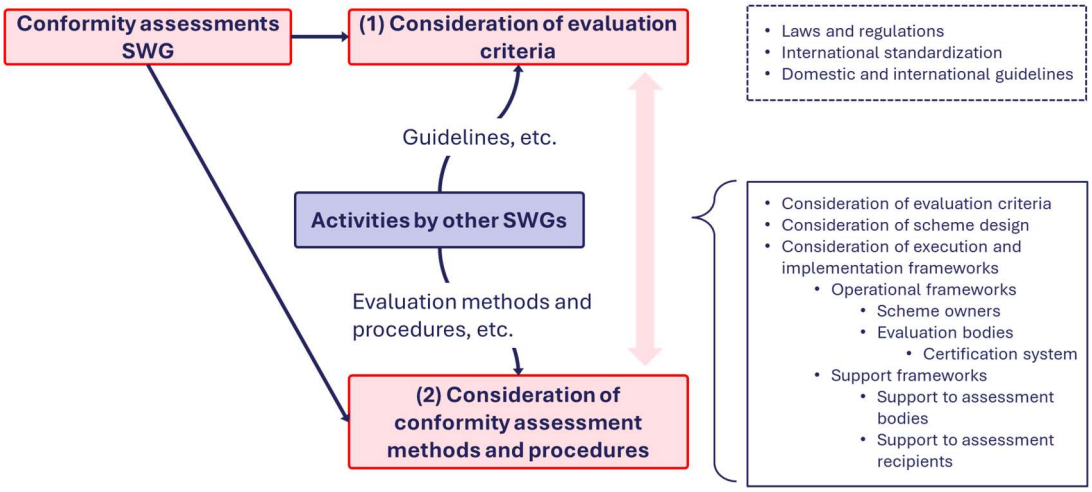

Figure 3-10: Approach to establishing conformity assessment methods

3.4.3. Roadmap

In FY2025, the SWG will collaborate with sector-specific SWGs to conduct interviews about evaluation needs and the current conditions of assessment recipients. It will also form a SWG composed primarily of relevant organizations involved in conformity assessments, proceed with the analysis of existing conformity assessment methods, and consider the composition of future SWG members from assessment recipients side.

In the medium term, the SWG will work on the design of conformity assessment schemes to ensure comprehensive assurance and accountability, based on the international standardization discussions under the ISO/IEC JTC1 SC42, including the work on ISO/IEC 42007. The SWG will also begin the elaboration and trial implementation of a conformity assessment scheme for AI in Japan, and transition to a phase of verifying its effectiveness for social implementation.

In the future, the SWG will promote the development of guidelines aligned with international implementation models to ensure global interoperability, and work on their phased deployment, with the goal of establishing a practical, internationally interoperable framework (domestic/international cooperation). (Figure 3-11)

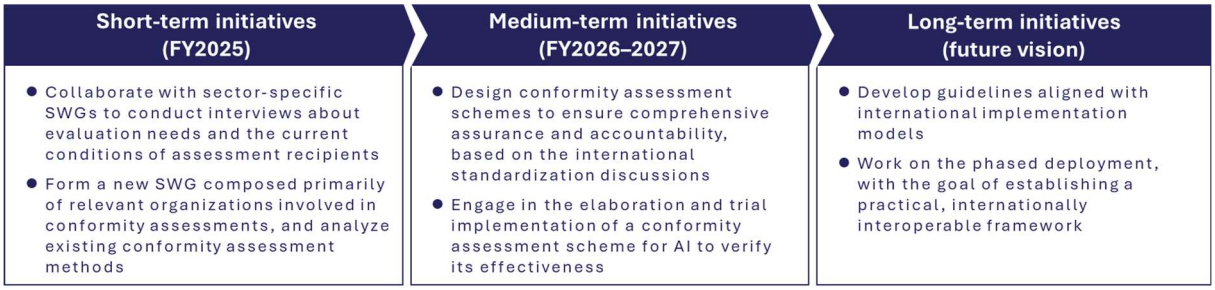

Figure 3-11: Roadmap of the Conformity Assessment SWG

4. Conclusion

AI utilization is becoming increasingly vital as a foundation for innovation that supports both the resolution of social issues and the transformation of industries. As its adoption accelerates, it is imperative to ensure that AI is introduced and utilized safely and with trustworthiness. This requires the establishment of a comprehensive process for appropriately identifying, evaluating, and mitigating risks with necessary measures.

Based on the directions outlined in this Vision Paper, it is expected that the effectiveness of AI safety evaluation will be enhanced through sector-specific, use case-driven demonstrations and the development of cross-sectoral evaluation frameworks aligned with international standards. The Business Demonstration WG aims to contribute to the sustainable growth of society and industry by providing a forum for advancing AI safety evaluations with the goal of facilitating the smooth implementation and utilization of AI technologies.

5. References

1. Ministry of Internal Affairs and Communications and Ministry of Economy, Trade and Industry. “AI Guidelines for Business,” Version 1.1, March 2025
   https://www.soumu.go.jp/main_sosiki/kenkyu/ai_network/02ryutsu20_04000019.html
   https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/20240419_report.html
2. Japan AI Safety Institute. “Guide to Evaluation Perspectives on AI Safety,” Version 1.1, March 2025
   https://aisi.go.jp/effort/effort_framework/guide_to_evaluation_perspective_on_ai_safety/
3. Japan AI Safety Institute. “Guide to Red Teaming Methodology on AI Safety,” Version 1.1, March 2025
   https://aisi.go.jp/effort/effort_framework/guide_to_red_teaming_methodology_on_ai_safety/
4. Japan Digital Health Alliance (JaDHA). “Guide to the Utilization of Generative AI for Healthcare Providers,” Version 2.0, February 2025
   https://jadha.jp/news/news20250207.html
5. Financial Data Utilizing Association (FDUA). “FDUA Generative AI Guidelines,” Version 1.0, October 2024
   https://www.fdua.org/news/news_20240828
6. Japan AI Safety Institute. “Data Quality Management Guidebook,” Version 1.0, March 2025
   https://aisi.go.jp/effort/effort_information/250331_2/