（Draft） AI Guidelines for Business
Appendix
January 2024

Appendix . Introduction

Structure of the Appendix and Expectations for Readers

IN THE MAIN BODY OF THE GUIDELINES FOR AI BUSINESS OPERATORS (HEREINAFTER REFERRED TO AS "THIS BODY"), THE basic principles (=why) that should be kept in mind by "AI developers," "AI providers," and "AI users," the entities covered by these Guidelines, and the guidelines (=what) that should be implemented with regard to AI based on these principles are presented. In order to realize the GUIDELINE, each entity needs to decide on a specific approach and put it into practice.

Appendix 1 describes examples of AI systems and services, concrete examples of their utilization, examples of entity patterns, examples of benefits from AI for each industry and business, and examples of risks based on actual cases, which are assumed in this guideline. In addition, Appendix 2 includes information on actions to be taken by service providers to establish AI governance, through action goals and practical examples to deepen understanding.

In addition, Appendix 3 provides explanations of important issues for "AI developers," Appendix 4 for "AI providers," and Appendix 5 for "AI users. The A section provides supplementary explanations and specific methods for implementing the important issues for each entity listed in Parts 3 to 5 of this document, while the B section provides specific methods for the particularly important contents of the "common guidelines" listed in Part 2 of this document, which are not listed in Parts 3 to 5 of this document. B describes specific methods for the particularly important contents of the "Common Guiding Principles" described in Part 2 of this volume.

In addition, Appendix 6 lists points to keep in mind when referring to the "Guidelines for Agreements on the Use of AI and Data," which can be used as a reference when making agreements that handle data. (Appendixs 7 to 9 listed in "Figure 1. Structure of these Guidelines" are prepared separately from this document.)

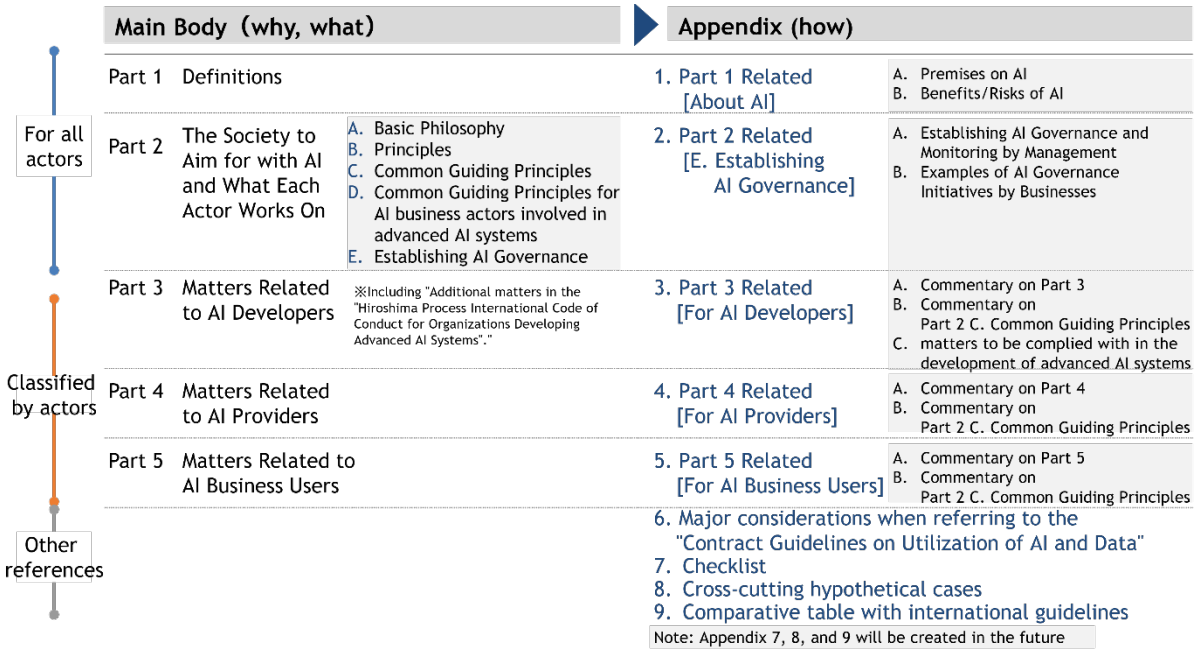

Fig. 1 . Structure of these Guidelines

It is assumed that by referring to Appendix 1 together while reviewing the descriptions in this volume and the Appendixs, it will be possible to understand AI and the benefits/risks of AI, which are the premises of the descriptions, in concrete terms, and to gain a deeper understanding of the contents of the descriptions. In addition, it is important for all entities of

"AI developers," "AI providers," and "AI users" to refer to Appendixs 1 and 2, as they can grasp the action goals for establishing AI governance in their own AI applications through specific practical examples by referring to Appendix 2.

Since Appendixs 3 through 5 are organized by entity, it is important for each entity to confirm the relevant contents, and to consider and implement measures based on the examples of practices that have been identified. It is expected that the contents of other entities should also be identified as much as possible, since referring to the contents of other entities together with the main report may lead to the consideration of measures such as risk reduction in the entire value chain.

In addition, it is also important to create and effectively utilize a checklist suited to the business content and situation of each business with reference to the checklist in Appendix 7 (attached material) in order to reliably promote the planning and implementation of initiatives to reap the benefits while minimizing the risks associated with AI. The checklist is formatted with 10 guidelines and important items to be checked in order to confirm whether the guidelines and important items described in Part 2C of this document are being implemented. The checklist is prepared on the assumption that each service provider will customize it according to their own circumstances and utilize it as necessary. It also includes a format that can be used by service providers involved in advanced AI systems as described in Part D of this document to confirm the implementation of important matters, and a format that can be used to confirm the establishment of AI governance as described in Appendix 2, thus contributing to the confirmation of implementation from initiatives to AI governance. It is expected that AI developers, AI providers, and AI users will cooperate with each other to consider optimal approaches, taking into account technological developments, changes in the external environment, etc., and it is assumed that this will help effective collaboration. The following is a brief overview of the project.

The Guidelines are based on the risk-based approach throughout this document and the Appendix, and it is expected that service providers will similarly identify what they need to focus on and what they do not need to focus on, and implement effective measures and AI governance structures. The Appendix is an example of a means to achieve the direction indicated in this volume, and does not comprehensively describe practices and explanations related to all the guidelines described in this volume. Therefore, it is not required to implement all of this appendix as described.

Explanation of Expressions in these Guidelines

The contents (items) listed in "Table 1: Items of importance for each entity in addition to the common guidelines" are identified and described in the same manner as in the main body of this document. The items are identified and described by the rule of [Subject - Guideline number].

- The entities are indicated by the initial letters of AI Developer, AI Provider, and AI Business User, and the numbers of the guidelines and descriptions are the same as the numbers in the table above.

(e.g., D-2) i. refers to key issues about learning appropriate data about the safety of AI developers

The "-" in the table does not mean that no action is required, as it is expected that each entity will take action based on the items described in Part 2, C. "Common Guiding Principles" of this volume.

Table 1. items that are important for each entity in addition to the "common guidelines".

	Part 2 . C. Common guidelines	In addition to the Common Guiding Principles, important issues for each entity		
		Part 3 . AI Developers (D)	Part 4 . AI Providers (P)	Part 5 . AI Users (U)
1) Man-centered	① Human Dignity and Personal Autonomy ② Attention to decision-making, emotional manipulation, etc. by AI ③ Countermeasures against false information, etc. ④ Ensure diversity and inclusion ⑤ user assistance ⑥ Ensure sustainability	-	-	-
2) Safety	① Consideration for human life, body, property, spirit and environment ② proper use ③ Appropriate learning	i. Learning of appropriate data ii. Human life, body and property, spirit and the environment. iii. Development that contributes to appropriate utilization	i. Risk measures that consider human life, body, property, spirit, and environment ii. Provision that contributes to appropriate use	i. Appropriate use with safety in mind
3) Fairness	① Consideration for bias in each component technology of the AI model Bias considerations included in each component technology of the AI model ② Intervention of human judgment	i. Consideration for bias in the data Consideration for bias in the data ii. Consideration for Bias in AI Model Algorithms, etc. Consideration for bias in AI model algorithms, etc.	i. The configuration of AI systems and services, and Consideration for bias in data Considerations	i. Consideration for bias in input data and prompts Consideration for bias in
4) Privacy Protection	① Protection of privacy in all AI systems and services	i. Learning of appropriate data (D-2) i. Restatement.)	i. Privacy Protection Implement mechanisms and measures to protect privacy ii. Measures against invasion of privacy	i. Improper Input of Personal Information and Measures Against Privacy Violations
5) Security Secured	① Security measures affecting AI systems and services ② Attention to the latest trends	i. Implement mechanisms for security measures ii. Attention to the latest trends	i. Implement mechanisms for security measures ii. Vulnerability Response	i. Implementation of security measures
6) Transparency	① Ensure Verifiability ② To relevant stakeholders Providing Information to Relevant Stakeholders ③ Reasonable and honest ④ To relevant stakeholders Accountability and Interpretability Improvement	i. Ensure Verifiability ii. To relevant stakeholders Providing Information to Relevant Stakeholders	i. System architecture, etc. Documentation ii. To relevant stakeholders Providing Information to Relevant Stakeholders	i. To relevant stakeholders Providing Information to Relevant Stakeholders
7) ACCOUNTERVIRABILITY	① Improved traceability ② Description of the status of compliance with the "Common Guiding Principles". Explanation ③ Identification of Responsible Persons ④ Distribution of responsibility among the parties involved ⑤ Specific Responses to Stakeholders Responses to Stakeholders ⑥ documentation	i. Common Guidelines for AI Providers Explanation of response status ii. Documentation of development-related information	i. Common Guidelines" for AI Users Explanation of response status ii. Documentation of terms of service, etc.	i. To relevant stakeholders Explanation ii. Use of the documents provided and Compliance with the terms and conditions
8) Education and Literacy	① Ensuring AI literacy ② Education and reskilling	-	-	-

	③ To Stakeholders Follow-up			
9) Ensuring fair competition	-	-	-	-
10) Innovation	① Promote open innovation, etc. ② Attention to interconnectivity and interoperability ③ Provide appropriate information	i. Contribution to the creation of innovation opportunities Contribution to the Creation of Innovation Opportunities	-	-

Appendix 1. Related to Part 1

A. Assumptions about AI

AI Learning and Use Flow

Generally speaking, AI is a system that builds an AI model based on data through a preliminary learning process, uses the model to make inferences and predictions, and outputs the results. In addition to conventional AI that uses AI models based on specific numerical data, images, and other data, these guidelines also cover generative AI that learns large amounts of text, images, and information posted on the Internet. In some cases, data obtained as output is used as input for relearning, and the output of one AI model may be used as training data for another AI model, or another AI model may be created from the original AI model (see "Figure 2. Example of AI Learning and Use Flow").

Appendix 1. Related to Part 1

AI Learning and Use Flow

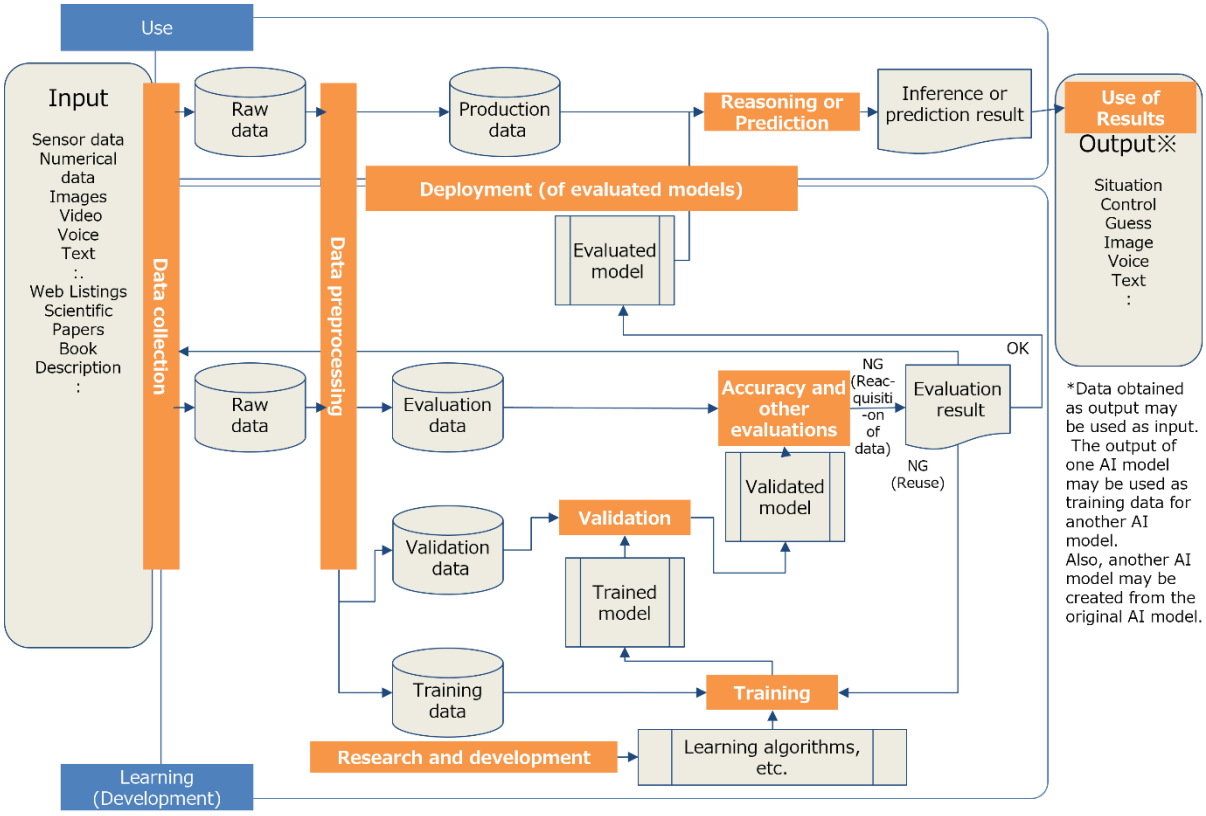

Figure 2: Example of AI learning and use flow

AI System Overview

A system in which software with AI functions is incorporated is treated as an AI system, which outputs via actuators and information terminals in response to inputs such as sensor data and text. In the Appendix, actuators are used as a generic term for devices that output images, sounds, text, and guessed results, in addition to driving devices such as electric motors and engines, and physical devices that are controlled by their operation.

In each phase of AI development, provision, and use, means such as fine tuning, transfer learning, reinforcement learning, and In-Context Learning (prompt engineering, memory, RAG: Retrieval-Augmented Generation, tool extension) In some cases, AI systems are improved and adjusted and updated through the use of AI systems (see "Figure 3. AI System Overview").

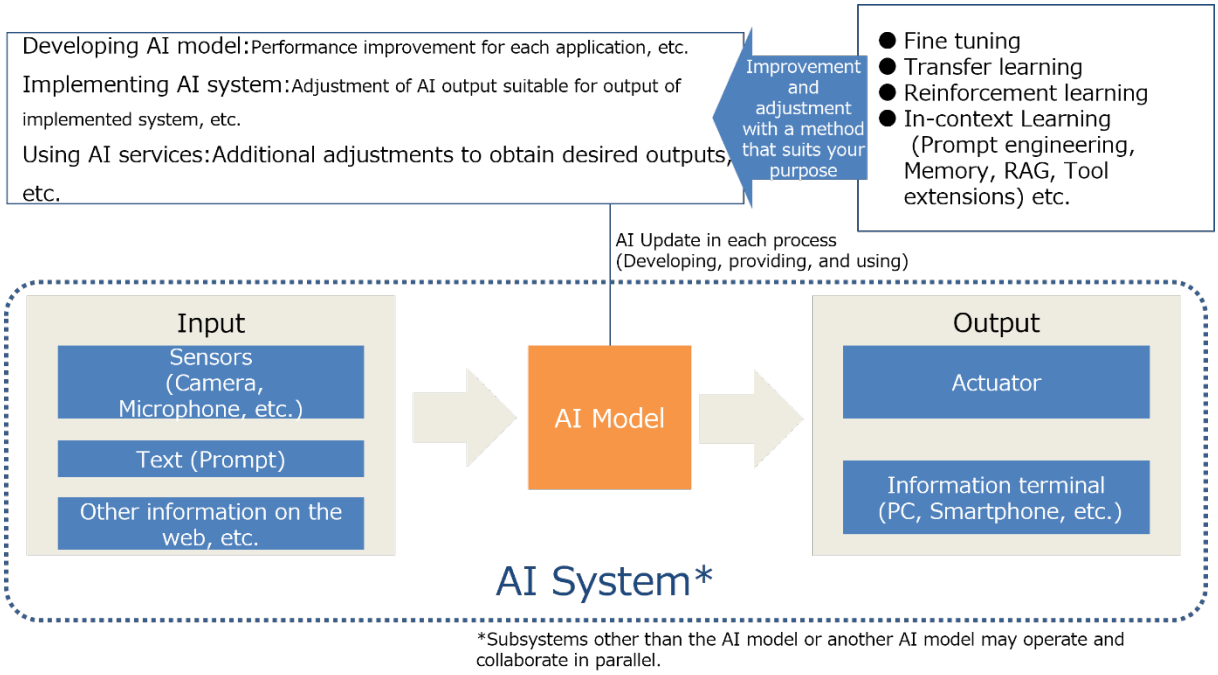

Figure 2 . AI System Overview

Value chain of AI from development to use

AI models are constructed by AI developers using the collected data, and AI systems are built by AI providers by incorporating the AI models into existing/new systems. The constructed AI system and AI services by the system are provided to AI users for their use (see "Figure 4. Subject's Response in the General AI Utilization Flow").

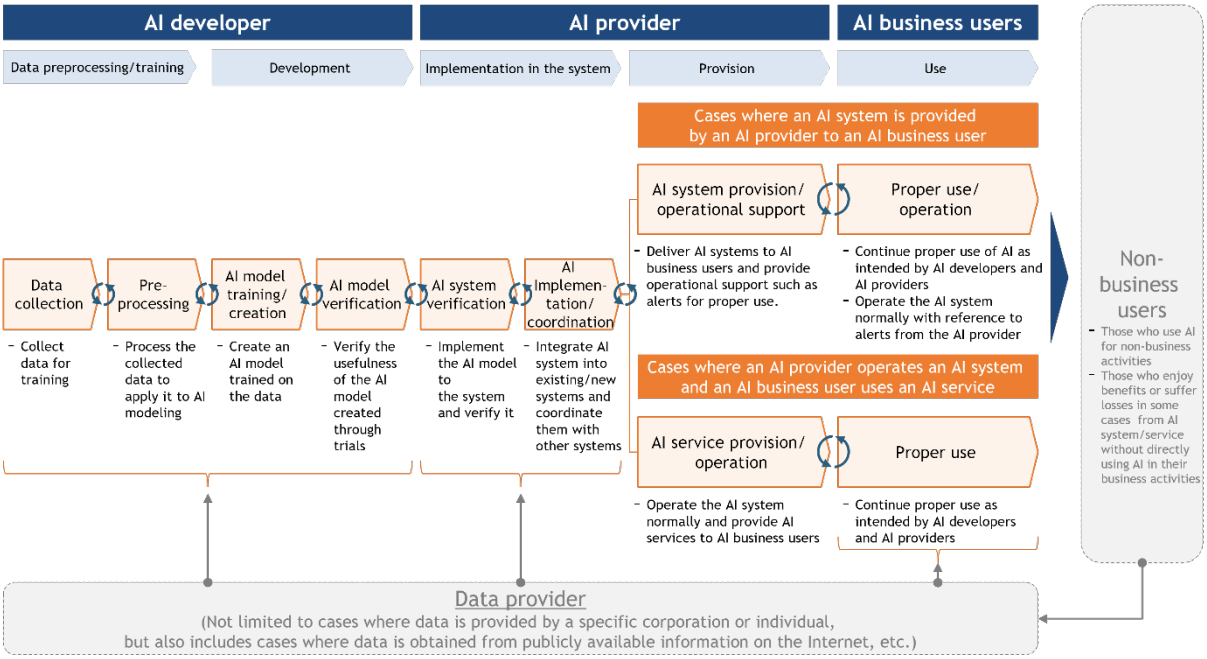

Figure 4 . Response of entities in the general AI application process

Examples of AI Systems and Services

Typical AI systems and services are listed in "Table 2.

Table 2. Examples of AI systems and services[1]

Case Name	Utilization AI	summary	AI Developer	AI provider	AI user	Out of Business user
Recruitment A I	Text analysis	This **AI service is used as reference information by the** human resource recruitment departments at each of the global companies in the Company A group **when judging the screening of entry sheet documents.** The AI development department of Company A receives past entry sheet data and results of pass/fail decisions (decisions on job offers) from Company A's human resource recruitment department (including overseas group companies), the AI user, and uses machine learning (classification model) to create an AI model to support pass/fail decisions.	Company A (Development Division)	Company A (Systems Division, Human Resource Development Division)	Company A Group (Recruitment department)	applicant for employment
unmanned convenience store	Image Analysis	Company J, which operates a nationwide chain of convenience stores, provides an **image recognition AI-based unmanned convenience store service (a convenience store where customers simply take an item from the store and the AI calculates the price, allowing them to make a lump-sum payment using electronic money or other means when they leave the store)**. This AI service is equipped with an AI system developed by Company X for unmanned convenience stores.	Company X	Company J (AI System Development Department, Convenience Store Division)	Convenience store	Convenience store User
Cancer Diagnosis AI	Text and image analysis	It uses multimodal learning and **takes "information related to the person's medical history, genetics, etc. (Data 1)" and "endoscope images (Data 2)" and highlights areas with a high possibility of cancer in real time during the endoscopic examination**. The physician observes the output images to determine if there is a possibility of cancer. Company A is developing AI while providing a cancer diagnosis AI system to medical institutions.	Company A (AI Development Division)	Company A (Medical IT Services Division)	Medical institutions (Systems Division, Gastroenterology)	patient being examined
Defective Product Detection AI	Image Analysis	**This is an "inspection system for finished products" using deep learning image generation and recognition models. In the** past, finished products (industrial parts) were inspected visually, which required a large amount of labor cost. The system **identifies defects in the appearance of finished products (industrial parts) produced at A Industry's factory**. Since the number of defects identified in the factory is extremely small in relation to the total number of finished products normally shipped, "AI models that generate images different from finished products" and "AI models that can correctly identify normal products" are utilized.	Company B (Manufacturing Solutions)	A Industry (Manufacturing Control Dept.)	A Industry (Production line @ factory)	-

[1] Excerpted from "Cases Listed in Risk Chain Model," Center for Future Vision Research, The University of Tokyo. AI developers, AI providers, AI users, and off-business users are listed in accordance with the entity organization of this guideline.

Appendix 1. Related to Part 1
Examples of AI Systems and Services

		The deep learning model is being developed by Company B, a contractor.				
Power line inspection AI	Image Analysis	**This is a "diagnostic service for overhead power lines" using image analysis technology based on deep learning. The service performs image analysis to inspect power lines and automatically detects abnormalities**. However, for power lines in mountainous areas and other environments where it is not easy to visually check the lines, it is necessary for skilled maintenance personnel to visually check the lines using slow playback of video images taken by helicopters, which requires a long time to perform. This took a long time. Against this backdrop, Company P decided to automate the process of identifying abnormalities in power lines and preparing reports by introducing image recognition AI from Company X. The images are taken by drone or helicopter. Although no decisions are made in real time, the image recognition AI identifies abnormalities and prepares a report as soon as the photographing work is completed.	Company X (AI Development Dept.)	Company P (Systems Division, Power Service and Maintenance Department)	Company P (In charge of maintenance)	-
Smart Appliance Optimization AI	Sensors Data Analysis	**AI models analyze environmental information, user behavior, etc., to optimize smart home appliances Company A's AI service** acquires sensor information (user location and condition, temperature, humidity, illumination, CO2 concentration), open data (weather information), and user feedback (stress, comfort level opinions, etc.) mounted in the space. The AI model analyzes this information and automatically controls smart home appliances (smart refrigerators (food management, recipe suggestions, etc.), air conditioning, floor heating, air purifiers, robot vacuum cleaners, ventilation systems, etc.).	Company A (AI Development Dept.)	Company A (Appliance Division)	-	consumer
Interactive AI in-house implementation	text generation	Company B's employees can enter prompts (instructions or questions) to the interactive AI **to** receive answers. The service is used for all purposes and applications within the company, including questioning, programming, document generation, translation, and summarization, and contributes to improved work productivity. Using Company A's cloud platform and generated AI model, Company B's group company, Company C, implements an AI assistant service and provides it to Company B's group employees (including Company C).	Company A	Company B Group Company C	Company B Group Employees (including Company C)	-

Patterns of AI Businesses

The value chain of AI used in business includes Pattern 1, in which AI is used by AI users and benefits are provided to non-business users[2] in addition to AI users, Pattern 2, in which AI users use and benefit from AI systems and services provided by AI providers and non-business users use and benefit from them. Pattern 3 (see "Figure 5. Patterns of AI providers").

In Pattern 1, the AI system (service) is not provided to non-working users, only the benefits.

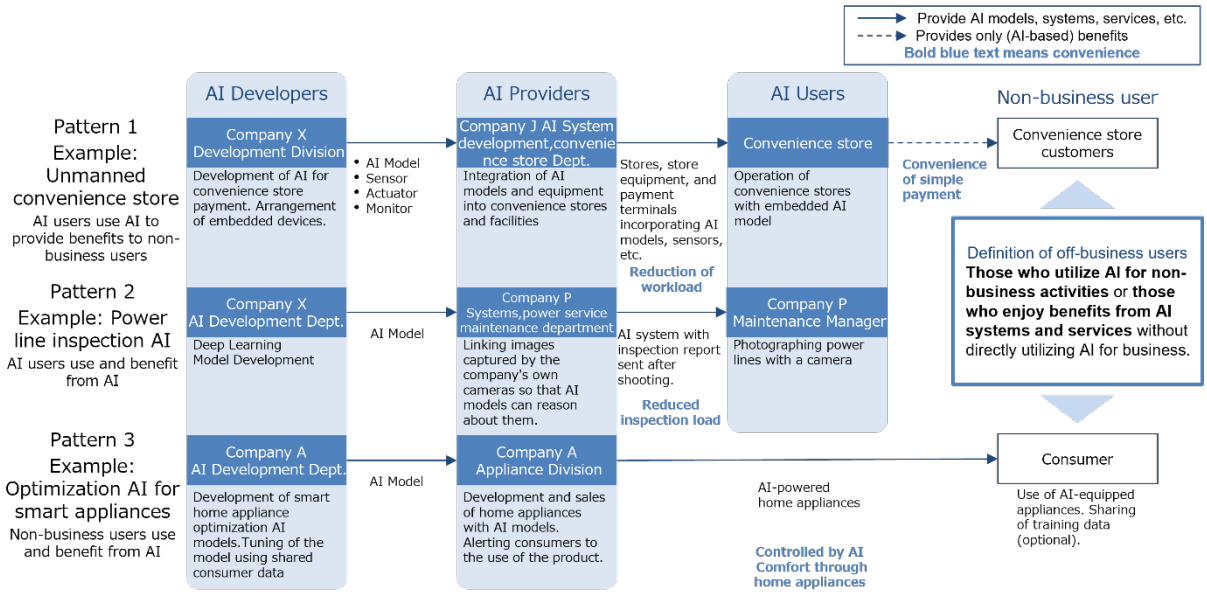

Figure 3 . Pattern of AI providers

[2] Those who utilize AI for non-business activities or those who benefit, or in some cases, suffer losses, from AI systems and services without directly utilizing AI for business (as defined in the main body of these Guidelines).

About Data Providers

In each phase of AI development, provision, and use, data are utilized to train AI models and use AI. In some cases, when building and using AI models with data, AI developers, AI providers, and AI users themselves use data held by themselves and do not use external data. On the other hand, there are cases where external data obtained from unmanageable data sources, such as data provided by specific corporations or individuals, or data obtained from people, sensors, or systems whose confidentiality, integrity, and availability are unmanageable, are used. Since it is extremely difficult to apply guidelines and important matters to uncontrollable data sources, this guideline describes the treatment of data concerning AI developers, AI providers, and AI users who fall under those who are provided with or obtain data, and are outside the scope of this guideline. Important matters for data providers are not described (see "Figure 6. How data should be provided").

However, when exchanging data with a specific corporation or individual, it is important to refer to Appendix 6 and the "Contractual Guidelines for AI and Data Use" referred to therein, and to proceed with the use of data upon agreement and contract between both the party to whom the data is provided and the party providing the data (data provider).

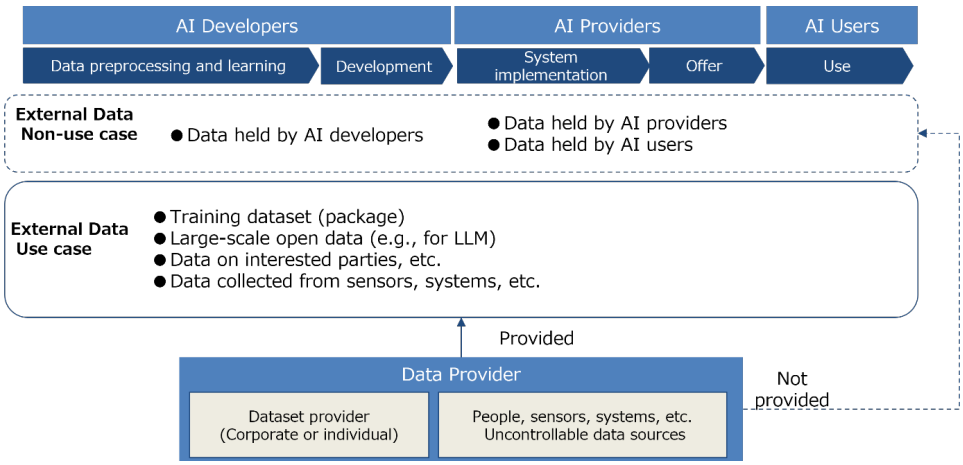

Figure 4 . The state of data provision

B.Benefits/risks from AI

While AI brings benefits, such as creating new businesses, adding value to existing businesses, and improving productivity, there are also risks.

This risk is expected to be controlled as much as possible. On the other hand, taking excessive risk countermeasures may increase costs and thus impede the benefits that can be obtained from AI utilization, so it is important to take a risk-based approach in which the degree of risk countermeasures corresponds to the nature and probability of the risk.

Benefits from AI

The benefits from the use of AI are manifold and continue to grow as the technology advances.

AI can be used to create value in each entity. As a result, the following can be expected

- Reduction of operating costs
- Create new products and services that accelerate innovation in existing businesses
- Organizational Transformation

In addition, applications in various fields (agriculture, education, medicine, manufacturing, transportation, etc.) and various deployment models (cloud services, on-premise systems, cyber-physical systems, etc.) are possible.[3]

Examples of Benefits

Figure 7 below shows just a few examples of the benefits of AI in corporate activities, but the benefits can be realized across the entire spectrum of corporate activities.

	Development	Marketing	Sales	Logistics Distribution	Customer relations	Judicial affairs	Finance	Human resources
From the past existing examples of benefits	Code verification, Documentation Automation	Automatic delivery of advertising e-mails	After receiving an order automatic dispatch of response e-mails, etc.	Demand forecasting production and inventory optimization	Automatic response by chatbot	Translation	Financial statements automatic creation	Automation of payroll calculation, etc.
With Generative AI further improvement	Extraction and verification of similar codes and data	Data-driven personalized advertising	Sales forecast (by channel and by need)	Delivery route optimization	Creation of FAQs based on past inquiries	Legal writing review	Past performance based on past results future projections fraud detection	Demand matching for human resources based on work history, etc.
Generative AI specific examples of benefits	Training data generation, Coding assistance, Brainstorming of new products	Automatic creation of sales promotion (marketing materials, catch copy, etc.)	Sales talks crypto's automatic creation	Logistics conditions egotiations assistance	Response details Automatic generation, Summary	In accordance with the regulations based on automatic generation of draft contracts	Responding to internal inquiries in context	Responding to HR interviews in context

Figure 7. Examples of Benefits from AI in Business Activities

For example, in the field of logistics, AI is used to automate delivery by robots and optimize the value chain through demand forecasting, and in the field of human resources, AI is used to automate payroll calculation and match human resource demand based on work history. AI is being used to streamline and optimize operations in a wide range of applications.

[3] ISO/IEC TR 24030 contains an extensive collection of use cases covering these areas and deployment models.

Outside of the corporate sector, there is also automation of administrative procedures, work support systems on farms using sensors and image information, and applications in the medical field using medical history and other information.

Furthermore, in the B2C realm, a wide variety of services are being developed, including chatbots, automated driving, search systems, and voice assistants.

Possibilities with Generative AI

In addition to the above, generative AI has emerged most recently. Generative AI is likely to trigger a turnaround for Japanese companies that have lagged behind in DX.

Japanese companies are characterized by their accumulation of high-quality operational technology (OT) data and their meticulous services and operations. When trying to realize these by utilizing conventional AI, a lot of man-hours and expertise are required to integrate data interfaces, prepare a large amount of data, create scenarios and cases assuming many patterns, and develop them based on the data, in order to utilize OT data across organizations and industries and to use AI in these services and tasks. In the past, the integration of data interfaces, the preparation of large amounts of data, the creation of scenarios and cases based on many patterns, and the development based on these scenarios and cases, required many man-hours and expertise. By utilizing generative AI, these scenarios and cases themselves can be automated (self-supervised learning), and this will promote the use of AI by a wide range of companies. In fact, there are examples of retailers using generative AI to create answers and materials for their call centers and sales representatives, thereby increasing their productivity. In addition, the system can generate multiple patterns of responses and materials by referencing internal data in response to inquiries and customer requests.

In order to survive the fierce global competition, companies are expected to have a correct understanding of the benefits they can enjoy, explore the possibilities, and take a proactive stance, for example, by reviewing their digital strategies in a way that proactively incorporates generative AI.

Risks with AI

While the benefits are expanding, the risks they create are also increasing with the expansion of use and the rise of new technologies. In particular, with the spread of generative AI, risks such as the generation and dissemination of false information and misinformation are diversifying and increasing, and there are growing calls for respect for intellectual property rights.

Specifically, the following cases have arisen:[4] . Note that the risks discussed here are representative and do not cover all risks of AI, but include cases based on assumptions, and are expected to be recognized as examples only. Therefore, the existence of this risk does not immediately preclude the development, provision, or use of AI.[5] Rather, it is expected that the development, provision, and use of AI will enhance competitiveness, create value, and eventually lead to innovation through proactive development, provision, and use of AI after recognizing the risk and considering the balance between the acceptability and benefits of the risk.

Note that risks are not disadvantages that occur to each business, but risks that occur to stakeholders[6] and society as a whole are also subject to consideration.

[4] For international case studies, see Partnership on AI, "The AI Incident Database (AIID)," with over 2,000 reports, http://incidentdatabase.ai が参考と The AIID is a database of more than 2,000 reports. For details, see "Column 1: Sharing Incidents" below. Note that the parentheses in each case indicate the corresponding "common guidelines" in this volume.
[5] Laws and regulations of other countries should also be noted. For example, the EU's "AI Act" describes AI systems that may pose a direct threat to human life and basic human rights (e.g., manipulation of the subconscious mind (therapeutic purposes are not covered)), social rating by the government, voice assistants that encourage dangerous behavior, etc., as "unacceptable risks" and discussions are ongoing (as of November 2020). The discussion is ongoing (as of November 2023).
[6] All entities that may be directly or indirectly affected by the use of AI, including AI developers, AI providers, AI users, and third parties other than non-business users (same hereafter).

Biased or discriminatory results output (person-centered, fairness)

- An IT company developed its own AI recruiting system, but discovered a flaw in the machine learning aspect of the system, which discriminated against women. The reason for this is said to be that the AI recognized that hiring men was preferable because most of the applicants in the resumes used for training over the past 10 years were men. The company in question attempted to improve the program so that it did not discriminate against women, but ended up dropping the operation as it would create another discrimination.

Filter bubble and echo chamber phenomena (anthropocentric)

- The division of society through recommendations by social networking services and other means has become an issue. For example, there is a concern that AI users and off-the-job users may become extremists through the filter bubble in which they are surrounded only by the information they want to see and the echo chamber phenomenon in which they are surrounded by people who think the same way as they do.

Loss of diversity (anthropocentric)

- If society as a whole uses the same model in the same way, opinions and answers may converge according to the LLM, and diversity may be lost.

Inappropriate handling of personal information (privacy protection, person-centered)

- The use of personal information that lacks transparency has become an issue. For example, in a service that uses AI for recruiting personnel, when the possibility of withdrawal from the selection process or withdrawal of a job offer was provided by AI, the explanation to students and other job seekers was unclear, and there were no rules for providing information to a third party based on temporary consent. The service was discontinued.
- Political use of personal information is also seen as problematic. For example, an election assistance campaign was conducted using a "personality diagnostic application" provided to off-the-job users of a social networking service and personal information collected based on profile information to understand and work with individual personalities to target advertising to encourage voting behavior in favor of the client. Specifically, based on the data collected, a large number of articles were posted in favor of their own camp, categorizing groups as "more prone to impulsive anger and conspiracy theories than the average citizen," "neuroticism and dark triad characteristics," and so on. This practice was feared to be an intervention in the election campaign using personal information and distorting democracy, which is the foundation of the country (negative impact on democracy).

Violation of life, limb, and property (safety and fairness)

- For example, an AI could make an improper decision that could cause a self-driving car to cause an accident, resulting in serious damage to life and property. In such a scenario, the risk of a large-scale accident due to AI malfunction is a concern
- In triage, where prioritization is performed at the time of an incident, ethical bias in the AI's ranking may cause a loss of fairness. When used in medical triage, discriminatory medical decisions may be made for certain groups of people, which may pose a threat to their lives.

Data Contamination Attacks (Secured)

- Risks include unauthorized data contamination of training data that can lead to performance degradation and misclassification when AI training is conducted, cyber attacks targeting the application itself when the service is operated, and attacks through AI inference results and prompts that are instructions to the AI. For example, a chatbot repeatedly uttered hate speech due to systematic learning of racist questions by a malicious group

Black boxing, demand for explanation of decisions (transparency, accountability)

- Problems have also arisen due to the black box nature of AI decisions. A report spread on social networking sites that a credit card offered to men and women with the same annual income had a lower credit limit for women than for men. In response to this problem, the financial authorities conducted an investigation and asked the company that provided the credit card to prove the validity of its algorithm. However, the companies were unable to explain the specific function and operation of the algorithm

Energy Use and Environmental Impact (Anthropocentric)

- The growing use of AI is also increasing the demand for computing resources, resulting in more data centers and increased energy use, and it has been suggested that the carbon dioxide emissions from the large amount of electricity used in AI development are dozens of times greater than the annual emissions of the American population.[7] However, it should not be forgotten that AI can also contribute to the environment, for example, by introducing AI into energy management, which will also enable more efficient use of electricity.

In addition, the risks manifested in the generated AI include the following.

Leakage of confidential information (ensuring security, education and literacy)

- In the use of AI, there is a risk that personal or confidential information may be input as a prompt and leaked through output from the AI. For example, in the use of AI services, there have been cases where employees have input source code that corresponds to confidential information into interactive generative AI for non-business users for business use. Since the barriers to the use of generative AI services have been lowered, there is a risk that employees may use non-work-related generative AI for high-risk purposes outside of corporate control, especially if corporate rules and regulations are not in place. However, there are interactive AI systems with enterprise-grade security features that are designed for business use. Companies are encouraged to use such services and applications instead, especially when processing sensitive information.

Abuse (safety, education and literacy)

- The use of AI for fraudulent purposes is also a growing problem. Among them, scams using AI-synthesized voices are on the rise. A woman received a call from her daughter asking for help and a ransom of $1 million, but it turned out that the voice was generated using AI and the call was a fraudulent call disguised as a kidnapping.

Halcination (safety, education and literacy)

- There have been lawsuits against AI developers and providers regarding "halcyonation," in which a generative AI responds plausibly to something that is not true. A TV program participant discovered that a generated AI was spreading false information that he was being sued for embezzlement of money. He even filed a lawsuit against the developer/provider of the AI for defamation, claiming that the AI had created a false complaint against him.

Belief in false information and misinformation (person-centered, education and literacy)

- The risk may be in taking advantage of misinformation generated by the generated AI. For example, a U.S. attorney's use of generated AI to prepare documents in an ongoing civil lawsuit resulted in a problem when he cited a precedent that did not exist.
- Deepfakes are being misused in a number of countries. Overseas, information manipulation and public opinion manipulation using fake images and videos have occurred. In one case,

[7] Stanford University, "AI Index Report 2023 - Artificial Intelligence Index," https://aiindex.stanford. edu/report/#individual-chapters

a fake image created with a generative AI claiming "an explosion occurred near the Pentagon" spread quickly on SNS and the Internet. Fake accounts posing as some foreign media and major financial media also spread this information, leading to a temporary drop of more than $100 in the average stock price. There have also been cases of corporate accounts spreading false information about incidents, accidents, disasters, etc.

Relationship with copyright (safety)

- Some stakeholders have raised arguments about the handling of intellectual property rights in the use of generative AI. Overseas, a number of artists have filed class action lawsuits claiming that when their works are trained by a generative AI to generate images, the AI sometimes generates images that resemble the works it has trained.[8]

Relationship with qualifications, etc. (safety)

- The risk of infringement of business law licenses and qualifications through the use of generated AI could also be considered. For example, if a generative AI answers legal or medical consultations, there could be infringement of business law licenses or qualifications, which could lead to legal problems. Attempts to avoid this risk could delay the adoption of generative AI throughout the industry and limit new services and efficiency gains.

Regeneration of bias (fairness)

- Because the generative AI creates answers based on existing information, if the situation in which the answers are believed continues, the biases contained in the existing information may be amplified and inequitable and discriminatory output may continue/expand. For example, if answers are created based on data from situations where gender discrimination exists, the risk of gender discrimination becoming entrenched increases as more people believe the answers

Thus, while the benefits of AI utilization are increasing due to technological development, risks that have emerged even with conventional AI are further increasing with the rise of generative AI. There are also risks that have newly emerged with the emergence of generative AI. In addition, as the barriers to use have been lowered for many generative AI services, there is a risk that they may be used in ways that involve unintended risks.

Generative AI is evolving day by day, and technologies and ideas for dealing with risks are also advancing day by day. However, the intrinsic risk of generative AI depends largely on its technical characteristics, and it is important to consider effective AI governance as a "better way to use" when considering countermeasures, so as not to end up with abstract discussions.

Furthermore, the risks of generative AI change with the external environment and technological trends, and are not "reproducible" and the causes of errors are difficult to identify. Therefore, socio-technical standardization, test validity, establishment of feedback loops, redefinition of legal and human rights risks, etc. are necessary to ensure contextually appropriate evidence.

In light of the above, there is a growing need to establish AI governance in order to enjoy the benefits of AI while controlling risks and enhancing competitiveness through the use of AI in business.

It is important to note that fear of risk is also a type of risk that prevents each entity from moving through "not utilizing AI until the risk is reduced to zero" or "pulling complete safeguards".

[8] In Japan, under Article 30-4 of the Copyright Act, in the learning and development stage, a copyrighted work may be used without the permission of the copyright holder to the extent deemed necessary, provided that the purpose is not to "enjoy" or cause others to "enjoy" the information analysis or other ideas or emotions expressed in the work. On the other hand, at the stage of creation and use, except for cases where the use is permitted under the Copyright Act, the judgment will be made based on dependence and similarity, as in the case of ordinary copyright infringement.

Appendix 2. "Part 2 E. Building AI Governance " Related

As described in Appendix 1.B. "Benefits/Risks of AI," in order to enjoy the benefits and control the risks of AI, it is important to establish AI governance to maximize the benefits of AI while managing AI-related risks at a level acceptable to stakeholders. In doing so, each entity is expected to apply the most appropriate solutions based on the constantly changing environment and goals, and continue to evaluate and review whether they are working properly.

Below are action goals as perspectives to be taken into account by each entity in establishing AI governance, as well as key points and examples of practices.

Of these, the behavioral goals are general and objective goals, and it is important that all entities involved in the development, provision, and use of AI systems and services that may pose a certain risk to society implement them (see "Table 3. List of Behavioral Goals" for the overall picture). On the other hand, which elements of the points of practice and practical examples assuming a hypothetical company are useful will differ depending on the individual and specific circumstances in which each entity is placed, and the objectives, methods, and evaluation targets of the AI systems and services developed, provided, and used by each entity. Therefore, it is left to each entity to decide on the points of practice and the adoption or rejection of examples of practice. Even in the case of adoption, it is expected that each entity will consider modification and selection according to its own circumstances.

In addition, it is expected that the AI governance system will be developed and operated in accordance with the requirements of stakeholders by coordinating with IT, privacy, and security governance, etc. within each entity, as well as between each entity throughout the value chain. In addition, in establishing AI governance, it is important to review mechanisms, rules, and systems in accordance with agile governance to speed up decision-making and operations while minimizing management man-hours. Promotion of appropriate delegation of authority is also expected to ensure proper AI governance and management, and efficient use of limited resources.

Table 3. List of Action Objectives

Classification.	behavioral goal
1. Environmental and Risk Analysis	1-1 Understanding benefits/risks 1-2 Understanding social acceptance of AI 1-3 Understanding your company's AI proficiency level
Goal setting	2-1 AI Governance Goal Setting
3. System Design	3-1 Evaluation of deviations from goals and mandatory response to deviations 3-2 Improvement of literacy of AI management personnel 3-3 Strengthening of AI management through cooperation among various entities and departments 3-4 Reduction of incident-related burden for users through prevention and early response
4. Operation	4-1 Ensure accountable status of AI management system operations 4-2 Ensure accountable status of individual AI system operations 4-3 Consider proactive disclosure of AI governance practices
5. Evaluation	5-1 Verification of the Functioning of the AI Management System 5-2 Consideration of External Stakeholders' Opinions
6. Re-analysis of environment and risk	6-1 Timely re-implementation of Action Objectives 1-1 through 1-3

A.Building AI Governance by Management and Monitoring

1. Environmental and Risk Analysis

Action Objective 1-1 [Understanding of benefits/risks]: Each entity, under the leadership of management, shall clarify the purpose of developing, providing, and using AI, and shall specifically understand the unintended risks as well as the benefits to be derived from AI in light of each entity's business, report these to management, share them with management, and update their understanding in a timely manner.

[Points of practice].

Each entity, under the leadership of management, will work to

- Clearly define the purpose of developing, providing, and using AI, such as creating value in business and solving social issues
- Specific understanding of "benefits" and "risks," including unintended ones, in a way that is tied to your business.
- In doing so, we pay attention to "risks" to be avoided and issues that span multiple entities, and ensure benefits and reduce risks throughout the value chain/risk chain.
- Establish a system for prompt reporting/sharing with management

Risks" include, specifically, the following, which could result in losses due to loss of reputation, fines for violations of laws and regulations, and liability for damages. For more details on risks, please refer to Appendix 1. "B. Benefits/Risks from AI".

- Risks common to AI in general
  - Biased or discriminatory output of results, filter bubble echo chambers, false information, improper handling of personal information, data contamination attacks, black boxing, leakage of sensitive data, misuse of AI systems and services, energy usage and environmental impact, re-generating bias, etc.
- Risks revealed by generative AI
  - Halcination, taking misinformation into account, relationship with copyright and other rights and qualifications, etc.
- Risks arising from organization and management
  - Inadequate recognition of the inclusion of AI in products and services, insufficient consideration of AI in governance, inappropriate or uneven use of AI due to lack of environmental awareness and planning, lack of organization of the relationship between humans and AI, such as segregation of work, etc.

In addition, issues that span multiple entities that are important for securing benefits and reducing risks throughout the value chain/risk chain include, for example, the following.

- Distribution of responsibility among or within entities
- Improvement of overall quality of AI systems and services
- Possibility of creating new value by interconnecting each AI system/service (System of Systems)
- Improve literacy of AI users and non-business users

As for reporting and sharing with management, it is expected that the most appropriate mechanism should be designed according to the characteristics of the company/organization, for example, the following methods may be considered.

- Establish internal bodies on AI governance that are accountable to the Board of Directors (AI Ethics Committee, AI Ethics Review Committee, etc.)
- Report on AI governance initiatives at Board of Directors meetings
- Document and circulate a document that lays out the benefits/risks of using AI for your company/organization.
- Reflection on the governance framework used in the company, etc.

[Practical example].
Practical example i: Identification of benefits and risks

It is important for each entity to consider not only benefits but also risks under the leadership of management (including implementation through management's own initiative, rather than entrusting it to the director in charge or the field site; the same applies below), to share the results of such consideration with management, and to update their understanding in a timely manner.

Although the benefits are considered to be known, we have reorganized the possible benefits of AI technology using comprehensive and exhaustive commentaries such as the "AI White Paper" compiled by the Information-technology Promotion Agency, Japan (IPA), .[9]

MHI has also investigated whether any incidents have occurred in the past in the same or similar functions or areas of the AI system or service that it intends to develop, provide, or use, or whether any specific potential for incidents to occur has been identified even if they have not occurred in the past. Incident information can be obtained from various documents and the Internet. Since we plan to develop, provide, and use our products only in Japan, we began by gathering information shared in Japan. In doing so, we referenced from the Consumer Affairs Agency's "AI Utilization Handbook - To use AI wisely"[10] . For example, one of the checkpoints listed was "AI may misrecognize your voice and give you wrong instructions or collect information about your usual conversations." is listed, but this is an expression of a potential incident from the perspective of a non-working user.

There is also a wide range of books on AI that refer to incidents and what could happen in the future. The Japan Deep Learning Association's (JDLA) G-Certificate also covers ethical matters, and information on incidents is available as part of this certification. The "Final Recommendations on Profiling" provides a clear explanation of some cases:[11] . Furthermore, while recognizing that social acceptance of AI systems and services can differ from country to country and region to region, we also referred to the incident database mentioned in "Column 1: Sharing Incidents" below. The analysis so far indicates that most of the incidents are related to the handling of personal information, fairness, and safety. The benefit/risk analysis of individual specific AI systems/services will be conducted during the deviation evaluation of Action Objective 3-1.

Practical example ii: Identifying risks using a framework when the scope of AI application is broad.

Given the diversity of the fields of AI systems and services that we develop, provide, and use, in addition to the practical examples i, we have included incidents that have had an impact on society and future issues that have been identified as having the potential to have an impact, in light of a general framework to get an overall picture of the issues, We have organized them broadly. We have developed and used our own framework, referring to the OECD Classification Framework[12] . In the Economic Context chapter of the OECD Classification Framework, which generally corresponds to environmental and risk analysis, the relationship between the OECD AI Principles and industry sectors, business The OECD's Economic Context chapter, which generally corresponds to environmental and risk analysis, presents a general framework from the perspective of the relationship between the OECD AI Principles and industry sectors, business applications, stakeholders, and scope of impacts. Keeping in mind that these classifications are only auxiliary tools for a broad understanding of risks, we are currently considering their reflection in our own framework. The benefit/risk analysis of individual specific AI systems/services will be conducted during the deviation assessment of Action Objective 3-1.

[9] AI White Paper Editorial Board "AI White Paper 2023" (May 2023)
[10] Consumer Affairs Agency, "Handbook for AI Utilization - To Use AI Wisely" (July 2020), https://www.caa.go.jp/policies/policy/consumer_policy/meeting_materials/review_meeting _004/ai_handbook.html
[11] Personal Data + Alpha Study Group, "Final Recommendations on Profiling" (April 2022), https://wp.shojihomu.co.jp/wp-content/uploads/2022/04/ef8280a7d908b3686f23842831dfa659. pdf
[12] OECD, "OECD Framework for the Classification of AI Systems: a tool for effective AI policies", https://oecd.ai/en/classification

Practical example iii: Understanding benefits/risks in collaboration with multiple entities when the scope of AI application is broad.

We are aware that the scope of development, provision, and use of AI systems and services is broad, and that incidents can have a significant impact on society. Therefore, we believe that the benefits/risks of AI can be analyzed more usefully by combining information obtained from our own direct involvement with the experience of other companies in the same industry and, in some cases, other industries, and we conduct such analysis in in-house study groups across the humanities and sciences. This analysis is then continued at a regular frequency so that AI governance goals can be reviewed in a timely manner, even before an incident occurs.

Example iv: Internal sharing of benefits/risks identified

The Company recognizes that the scope of development, provision, and use of AI systems and services is broad and that its impact will be felt throughout society. Therefore, we believe it is important to share the benefits and risks of AI within the company and have taken the following steps

First, the information obtained and the results of the analysis are summarized, and the documents are documented and circulated to the relevant parties within the company. This document is open to comments and feedback from relevant parties, and an active exchange of opinions takes place. Specifically, through internal study sessions and workshops, discussions are held with relevant department personnel and interested members of the company to incorporate opinions from different perspectives.

In addition, within the company, a dedicated person is designated to be responsible for AI governance initiatives and progress, and reports to the Board of Directors. This facilitates effective communication with management.

Through these internal sharing mechanisms, transparency is ensured, creating an environment in which the entire organization can maximize the benefits of AI while at the same time appropriately controlling risks.

Practical example v: Response to generative AI

Generative AI has also emerged in the recent past, and the company sees this as an opportunity for itself. In utilizing it for its own operations, the company is formulating its own in-house usage guidelines in line with the "Guidelines for the Use of Generated AI" published by the JDLA at[13] . At the same time, the company also confirms the information dissemination from the government, such as "Alerts, etc. on the Use of Generated AI Services" issued by the Personal Information Protection Commission. It is also essential to collect information on generated AI through news, SNS, and other media. Through these efforts, we strive to keep abreast of the latest trends in benefits and risks.

Column 1 Incident Sharing

There is much to be learned from past incidents about risks associated with the development and operation of AI systems; since AI systems are built inductively based on data sets and many of the risks are unintended, understanding past incidents is useful for reducing risks. Incident cases are generally obtained from news, papers, and other public information, but accessing the necessary information is not easy.

To address this accessibility challenge, the Partnership on AI released The AI Incident Database (AIID)[14] in November 2020. AIID lists over 2,000 incidents with URL links and provides an app for searching. In addition to Partnership on AI, the AI Incident Tracker is available on GitHub[15] . The OECD has also released the OECD AI Incidents Monitor (AIM)[16] . AIM monitors global news as incidents, with over 150,000 daily news articles provided by the

[13] Japan Deep Learning Association, "Guidelines for the Use of Generative AI Version 1.1" (October 2023), https://www.jdla.org/document/#ai-guideline
[14] Partnership on AI, "AI Incident Database," https://incidentdatabase.ai/
[15] jphall663, "awesome-machine-learning-interpretability", https://github.com/jphall663/awesome-machinelearning - interpretability/blob/master/README.md#ai-incident-tracker.
[16] OECD, "OECD AI Incidents Monitor (AIM)," https://oecd.ai/en/incidents

Event Registry, a news intelligence platform. English-language articles provided by the Event Registry, a news intelligence platform, are analyzed for publication.

On the other hand, it seems to be a challenge to maintain such a database on a sustainable basis, as most of AIID's incident cases are initial lists provided by academics, etc. As the provision of AI continues to increase, it is also a challenge to accumulate important information. It is also pointed out that it is not easy to actively collect incident cases and turn them into common property, since the near-misses of each company that have not become public information are in themselves important experiences and may become the intellectual property of each company.

Action Objective 1-2 [Understanding the social acceptance of AI]:
Under the leadership of management, each entity is expected to understand the current status of social acceptance of AI based on stakeholder opinions prior to full-scale development, provision, and use of AI. Even after the full-scale development, provision, and use of AI systems and services, they are expected to reconfirm the opinions of stakeholders in a timely manner in light of changes in the external environment.

[Points of practice].
Each entity, under the leadership of management, will work to

- Identify stakeholders
- After identifying and working to understand social acceptance, develop, provide, and use AI.
- Reaffirm stakeholder input in a timely manner as needed, taking into account the rapidly changing external environment, even after the start of the offering

Stakeholders should be identified after considering the benefits and risks to individuals, organizations, communities, society, and the global environment that the AI to be provided will have throughout its life cycle. Note that it is expected that the scope is likely to be larger than the assumed stakeholders. For example, OECD's classification framework lists the following as stakeholders.

- Persons belonging to each entity
- non-business user
- Business
- government body (agency)
- research institution
- Scientist/Researcher
- citizen's group
- Children and other vulnerable groups, etc.

To understand social acceptance, it is useful to refer to the following information

- Public documents, academic research, etc.
  - Surveys published by governments and think tanks
  - research thesis
  - Opinions from civil society on AI systems and services
  - Seminars and conferences on AI ethics and quality
- Latest News
  - Investigation of incident cases
  - Stakeholder response, including non-business users on social networking sites, blogs, bulletin boards, news reports, etc.

The external environment of an organization, for example, "ISO/IEC23894:2023"[17] lists the following

- Social, cultural, political, legal, regulatory, financial, technological, economic, and environmental factors
  - Relevant laws and regulations, including those related to AI
  - Guidelines on AI issued by government, civil society, academia, industry associations, etc.
  - Sector-specific guidelines, frameworks, etc.
- Factors and trends affecting organizational goals
  - Technological trends and advances in various fields of AI
  - Social and political implications of the introduction of AI systems, including their organization in social scientific guidelines
- Stakeholder relationships, perceptions, values, etc.
- Contractual relationships and commitments to them

[17] ISO, "ISO/IEC 23894:2023(Information technology-Artificial intelligence-Guidance on risk management)" (February 2023)

- Complexity of coordination and dependencies among AI systems, etc.

In addition, the following methods can be used to reconfirm stakeholder opinions

- Direct feedback from stakeholders
- Evaluation of the company's AI management system and operations by AI experts

[Practical example].
Example i: Understanding Social Acceptance

We have tried to understand the social acceptance of AI, taking as our first clue the questionnaires for non-business users published by the government, public organizations, and think tanks. For example, the Consumer Affairs Agency, in its "AI Working Group of the Study Group on Consumer Responses to Digitalization," conducted a survey on "1) consumers' understanding of AI, 2) consumers' expectations, issues, and intentions to use AI, 3) AI-provided services used by consumers (what kind of risks they face), 4) AI's The Company conducted a questionnaire survey on the extent to which consumers recognize and understand the risks associated with AI services, and published the results of the survey. Since we are also considering international expansion, we also referred to the questionnaire survey of non-business users outside of Japan. Furthermore, we also referred to the opinions of civic groups on AI systems and services.

The information on social acceptance obtained here will be used in the overall design of AI governance, and therefore, it is necessary to shave off the branches and leaves and extract the trunk information so that management can make decisions. While utilizing the information and analysis obtained in Action Objective 1-1, we will identify various AI systems and services as applications that are likely to not reach the level of social understanding by any explanation, applications that are likely to be socially understood by proactive and sufficient explanation, and applications that are likely to be socially understood by explanation as necessary. The social acceptance is organized on a risk basis by classifying uses according to the magnitude of risk, such as uses that are likely to be socially understood through active and sufficient explanation, and uses that are unlikely to pose a risk to users outside of business operations.

Practical example ii: Understanding social acceptance by utilizing external seminars, etc.

In addition to Practical Examples i, we actively send our staff to seminars and conferences on AI ethics and quality held by universities and industry associations. Recently, these seminars and other events are often held in the form of webinars, making it possible to obtain information more efficiently than before. It is also possible to keep abreast of international trends in AI ethics and quality by accessing international webinars.

Practical example iii: Understanding social acceptance through stakeholders

Although we have adopted the methods described in Practices i and ii, we understand that our stakeholders have relatively high expectations of our appropriate use of AI, given that we have developed, provided, and used AI systems and services on a full-scale and extensive basis. Therefore, under the leadership of management, the Company has switched to a policy of directly and proactively capturing stakeholder opinions rather than indirectly and passively.

Under this new policy, we have invited experts who are familiar with the situation of social acceptance of AI and hold regular meetings including external experts on AI governance. We use this meeting not only to obtain evaluation results of our AI management system and operations, but also to deepen our understanding of the environment in which we are placed, including the general social acceptance of AI. We also recognize that, compared to the general information obtained in Practice Examples i and ii, the information obtained at the meetings is more in-depth for our company, and is often not widely known. The information obtained at these meetings is then combined with the general information obtained in Examples i and ii, and analyzed in detail on a risk-based basis for social acceptance. The results of the analysis are organized at the management level of the conference body, and the management level reports the results to the management level (in charge of business execution).

Action Objective 1-3 [Understanding the company's AI proficiency level]:
Under the leadership of management, and based on the implementation of Action Goals 1-1 and 1-2, each entity, except in cases where the risk is judged to be negligible in light of the intended use of AI, the company's business domain and size, etc., should consider the following factors The company's AI proficiency level should be assessed and reassessed in a timely manner based on the number and experience of employees, including engineers, involved in the development, provision, and use of AI systems and services, and the degree of literacy of such employees with respect to AI technologies and ethics, among other factors. If possible, the company is expected to disclose the results to stakeholders to a reasonable extent. If the company decides that the risk is not significant and does not assess AI proficiency, it is expected to disclose to stakeholders the fact that it does not assess AI proficiency and the reasons why it does not assess AI proficiency.

[Points of practice].
Each entity, under the leadership of management, will work to

- Consider the necessity of AI proficiency assessment in light of each entity's business domain, size, etc.
- If deemed necessary, visualize AI's ability to respond to risks and assess AI proficiency (how well prepared it is required when developing, providing, and using AI systems and services)
  - Disclose the results to stakeholders, to the extent reasonably possible
- If not deemed necessary, disclose that fact to stakeholders, if possible and to the extent reasonably possible, along with the reasons for the decision

Successful implementation of AI systems and services can bring benefits to businesses, such as eliminating human resource shortages, improving productivity, and developing high value-added businesses. On the other hand, unrestrained business provision of AI systems and services may unintentionally impair fairness, raise safety issues, and entail other risks inherent in AI. Therefore, each entity is required to start AI implementation with an understanding of these risks, which may be called the negative aspects of AI implementation, and for this purpose, AI proficiency assessment is important.

In order to assess AI proficiency, the following guidelines may be useful. Note that any of the guidelines may be revised in light of changes in the environment, including progress in the use of generated AI, so it is expected that the latest status be checked.

- Guidelines published in Nippon Keidanren's "Toward the Realization of Society 5.0 for SDGs through the Application of AI" (June 2023).[18]
- Certification test conducted by the Japan Deep Learning Association
  - Generative AI Test[19]
  - JDLA Deep Learning For GENERAL (G Test)[20]
- NIST, "Artificial Intelligence Risk Management Framework" (AI RMF 1.0)[21]

[Practical example].
[Practical example i: Assessment of proficiency level using the guidelines published in "Toward the realization of Society 5.0 for SDGs through the use of AI" (June 2023)].
Under the leadership of senior management, the Company assesses and reassesses its AI proficiency level in a timely manner to ensure that stakeholders do not suffer significant damage as a result of the introduction of AI systems and services due to insufficient consideration of risks when developing, providing, and using AI systems and services, because

[18] Nippon Keidanren, "Using AI to Achieve Society 5.0 for SDGs" (June 2023), https://www.keidanren.or.jp/policy/2023/041.html
[19] Japan Deep Learning Association, "Generative AI Test," https://www.jdla.org/document/#ai-guideline JDLA https://www.jdla.org/certificate/generativea
[20] Japan Deep Learning Association, "What is the G-Test?", https://www.jdla.org/certificate/general/
[21] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," https://nvlpubs.nist.gov/nistpubs/ai/NIST. AI.100-1.pdf

they are only concerned about the benefits. The company assesses and reassesses its AI proficiency in a timely manner.

For the evaluation of AI proficiency, the guidelines published in "Toward the Realization of Society 5.0 for SDGs through AI Application" (June 2023) by Nippon Keidanren (Japan Business Federation) are used. The reason for this is to evaluate whether or not the magnitude of benefits/risks that the company's AI systems/services can provide to society and the breadth of relevant stakeholders[22] are commensurate with the company's AI proficiency level. The Company then uses the AI proficiency level to inform its overall AI governance considerations, including consideration of AI governance goals.

[Practical example ii: Assessment of proficiency using original indicators while referring to the guidelines published in "Toward the realization of Society 5.0 for SDGs through the use of AI" (June 2023)].

The Company has just begun its AI governance study efforts in earnest. Therefore, as an evaluation of AI proficiency, we have selected several items from the guidelines published in "Toward the Realization of Society 5.0 for SDGs through AI Application" (June 2023), while referring to the guidelines, and have created our own indicators suitable for our own AI governance. The company is planning to use the results of the evaluation using these indicators to spread the current AI governance system and mechanism within the company, and to use more items to measure AI proficiency in the future.

Practical example iii: Assessment of proficiency on generative AI

In order to incorporate the impact of the recent rise of generative AI, we are using the "Guidelines for the Use of Generative AI" published by JDLA ([23]) to evaluate proficiency levels, taking into account the elements of generative AI. We have also heard that the "Guidelines for AI-Ready" are scheduled to be updated to take into account the use of generated AI, and we plan to conduct another review based on the updated guidelines when they are released.

[22] Entities directly or indirectly involved in the use of AI through the use of AI, including AI developers, AI providers, AI users, and non-business users

[23] Japan Deep Learning Association, "Guidelines for the Use of Generative AI Version 1.1" (October 2023), https://www.jdla.org/document/#ai-guideline

Goal setting

Action Goal 2-1 [Set AI Governance Goals]:
Each entity, under the leadership of senior management, will consider whether or not to set its own AI governance goals (e.g., AI policy), taking into account the benefits/risks that AI systems/services may bring, public acceptance of the development, provision, and use of AI systems/services, and its own AI proficiency level, while also noting the importance of the process leading to the setting of AI governance goals. Consider whether or not to establish AI governance goals (e.g., AI policies), taking into account the importance of the process leading to the establishment of AI governance goals. It is also expected that the established goals will be disclosed to stakeholders. If AI Governance Goals are not set because the potential risks are not material, the fact that they are not set, along with the reasons for not setting them, is expected to be disclosed to stakeholders. If the "common guiding principles" in these Guidelines are judged to be sufficient, the "common guiding principles" may be used as a goal in place of the company's AI Governance Goals.
Even if you do not set a goal, you are expected to understand the importance of this guideline and implement actions related to Action Goals 3 through 5 as appropriate.

[Points of practice].
Each entity, under the leadership of management, will work to

- Consider whether to set "AI Governance Goals" for each entity
  - Flexible settings based on the size of each company and the risk of the AI handled.
- Set a goal if you deem it necessary
  - Disclose such goals to stakeholders to the extent reasonably possible
- If not deemed necessary, disclose that fact to stakeholders, if possible and to the extent reasonably possible, along with the reasons for the decision.

The following are possible components of the "AI Governance Goal," and representative examples can be found in various references

- The company's own action policy, consisting of items that correspond to the "common guidelines" described in this guideline (the term "AI policy" or other terminology may vary depending on each individual company).
- In addition to the items addressed in the "Common Guiding Principles," a privacy policy that outlines guidelines for the use of privacy-related data, etc.
- Policies to increase inclusiveness and other benefits from AI utilization
- Tolerance for Risk

In addition, at the stage of preparing the AI Governance Goals, which are intended to be disclosed externally, it would also be useful to privately establish a code of conduct for employees and disseminate it internally (especially to those in charge of the practice) to raise employee awareness.

The "Common Guiding Principles" described in these Guidelines can also be used as "AI Governance Goals," and it is expected that the contents of the "Common Guiding Principles" will be used as a reference when setting AI Governance Goals unique to each entity. When AI Governance Goals are organized based on the Common Guiding Principles, possible risks can also be organized by linking them to the Common Guiding Principles, which will enable risk assessment based on the Common Guiding Principles.

In establishing the "AI Governance Goals," the following items shall also be taken into consideration

- Consideration of AI-related goals such as "AI governance goals" and AI utilization objectives in line with management goals such as the raison d'etre, philosophy, and vision of each entity, so that there are no conflicts or contradictions among them.

Goal setting

- Management goals, such as the raison d'etre, philosophy, and vision of each entity, as well as AI-related goals consistent with these goals, should be communicated when the PDCA cycle based on the AI Governance Goals is incorporated into organizational management.
- To identify stakeholders, consider the impact expected by stakeholders and the risks that stakeholders are concerned about, and ensure that there is no conflict with this.

[Practical example].
[Practice i: AI Governance Goals not set
We have just started developing AI systems, and for the time being, we plan to deal only with AI systems and services for applications where the potential risk to society is negligible. Therefore, we have not set AI governance goals, but we will consider setting AI governance goals when we expand the scope of our business to applications where the potential risks are not so minor. Of course, we will document our consideration so that we can explain to our stakeholders why we do not set AI governance goals.

[Practical example ii: Setting AI governance goals for a small business].
Although we have just started developing AI systems, we have decided to start AI governance efforts because the risks associated with the AI systems we develop are not insignificant. However, it is difficult to set up a special person in charge of AI governance because the number of employees is small. Therefore, the management drew up a "company's AI development policy" in line with the company's management philosophy, and shared it with the staff in charge of AI development. Then, the management and employees shared actual incidents that had occurred, and brushed up the "company's approach to AI development" and coordinated the company-wide viewpoints. As a result, it was decided to set this as the company's AI governance goal, albeit on a single A4 page.

[Practical Example iii: Setting AI Governance Goals Involving Each Department].
The Company has a diverse business portfolio, and each division has a different approach to its involvement in AI technology. In addition, given that we have adopted a company system, each of which is independent of the others, it is not easy for each division to agree on a single AI governance goal. Therefore, at this point, we will respect the "common guiding principles" of these guidelines, and in parallel, we aim to increase understanding of AI ethics and quality by adding AI ethics and quality as part of company-wide training on AI. Furthermore, an AI consultation service has been established within the company to collect case studies from various departments. Although it may appear to be moving slowly externally, we believe that the process of agreeing on AI governance goals is worthwhile. In addition, it is possible to consider the necessity and content of AI governance goals for each department that develops, provides, and uses AI systems and services at the stage before setting AI governance goals for the entity as a whole.

[Practice iv: Setting AI Governance Goals with Stakeholder Involvement].
In addition to developing, providing, and using AI systems and services, the Company has extensive experience in supporting other entities to develop, provide, and use AI systems and services for applications where the potential risk is seen as not insignificant. Although no serious incidents have ever occurred from AI systems/services developed by the Company or provided to other companies, the Company understands that there are many applications of AI systems/services provided by the Company for which social acceptance has not yet been established. Therefore, we have established and publicly announced our AI Governance Goals in order to enhance communication with our stakeholders. Since stakeholders understand the Company's policy, it is evaluated that the personnel in charge of developing AI systems and services and stakeholders can share the same basic stance toward AI technology, and communication has been facilitated.

3. System design (construction of AI management system)

Action Objective 3-1 [Assessment of deviations from goals and mandatory response to deviations]:
Each entity, under the leadership of management, should identify deviations of its AI from the AI governance goals, evaluate the impact of such deviations, and, if risks are recognized, determine whether the acceptance of such deviations is reasonable, taking into account their magnitude, scope, frequency of occurrence, etc. If the acceptance is not reasonable, the entity should be encouraged to reconsider the development, provision, and use of AI systems and services. If acceptance is not deemed reasonable, a process to encourage reconsideration of the development, provision, and use of AI is expected to be incorporated into the overall AI management system and at appropriate stages, such as the design stage of AI systems and services, the development stage, before the start of use, and after the start of use. It is important for management to formulate basic policies and other policies regarding the reconsideration process, and for the management layer to materialize this process. It is also expected that those who are not directly involved in the development, provision, and use of the target AI should be included in the evaluation of deviations from the AI governance goals. It is not appropriate to arbitrarily disallow the development, provision, and use of AI solely because of the existence of deviations. Therefore, the deviation assessment is only a step to evaluate the risk and a trigger for improvement.

[Points of practice].
Each entity, under the leadership of management, will work to

- Identify and assess deviations of current AI systems and services from "AI Governance Goals".
- If the risk is recognized, determine whether or not the acceptance of the risk is reasonable.
- Process for reconsidering/reconsidering how development, provision, and use of[24] should be if acceptance is not deemed reasonable, and incorporation into the decision-making process at appropriate stages of development, provision, and use and in organizations within each entity.
- The above is implemented on an ongoing basis, with management taking leadership and responsibility for decision-making, and the operational layer taking specifics.
  - Recognize and address the fact that the responsibility for establishing AI governance, organizational management and project management frameworks is equally as heavy as the operational responsibility.
- Share the determined discrepancy evaluation items within each entity in order to foster awareness within each entity.
  - Conduct deviation assessments in collaboration with each entity, depending on the content of the AI to be provided.

In some cases, the deviation evaluation process (to measure how well the AI system functions as designed and how accurately it can perform tasks such as forecasting and inference) is expected to be developed by referring to the following documents in accordance with the company's situation and the level of risk of the AI system/service, while utilizing the knowledge of outside experts. The following is an example of the process that is expected to be used

- Standard deviation evaluation process for each industry as described in Action Objective 3-1-1
- NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)."
- OECD, "FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS."
- Alan Turing Institute, Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems

[24] In development terms, it is generally referred to as CI (Continuous Integration) or CD (Continuous Delivery).

ISO/IEC 42001 states that it is important to ensure the integration of AI management systems and organizational business processes under the leadership of senior management.

- Ensure adequate resources for the AI management system.
- Communicate within each entity the importance of effective AI management and the importance of conforming to the requirements of the AI management system.
- Ensure that the AI management system achieves the intended AI governance goals
- Direct and support personnel who contribute to the effectiveness of the AI management system.
- Facilitate continuous improvement.
- Provide support to other relevant entities and exercise leadership in their areas of responsibility.

[Practical example].
[Practical example i: Deviation evaluation process for small businesses

Since we are a small company, the technical officers and development staff are close to each other, and the number of projects is not so large, the technical officers are fully aware of all the projects. The director in charge of technology establishes viewpoints for evaluating deviations from the "common guidelines" of these guidelines, and instructs development staff to identify deviations for each viewpoint, evaluate the impact caused by the deviations, and report to the director in charge of technology on all AI system development projects at the earliest practicable stage. The technical officer is then instructed to report back to the development director. If there is a risk, the officer in charge of technology shall reevaluate the impact caused by the deviation based on the report of the person in charge of development, and if there is a risk, the officer in charge of technology shall judge whether the acceptance of the risk is reasonable or not, and if the acceptance is not reasonable, the officer in charge of technology shall reconsider how AI should be provided. If the acceptance is not deemed reasonable, the company will reconsider how to provide AI.

In operating this process, in accordance with Action Objective 3-1-1, we refer to the standard deviation assessment in our industry and these guidelines.

Practical example ii: Deviation evaluation process for each project of a business with many divisions.

The Company, which has numerous divisions, has determined an AI Governance Officer and has established a committee on AI Governance under this officer. This committee, which consists of persons other than those in charge of projects for development, provision, and use of specific AI systems and services, is tasked with conducting, on a project-by-project basis, an evaluation of deviations from the AI Policy established by the Company based on the "common guidelines" of these Guidelines. Specifically, the Company shall prepare an evaluation list based on the AI Policy, identify deviations in the development, provision, and use of AI systems and services using the evaluation list, evaluate the impact caused by the deviations, determine whether the acceptance of the risk, if any, is reasonable, and if the acceptance is not reasonable, determine whether the development, provision, and use of AI systems and services should be implemented in accordance with the AI Policy. If acceptance is not deemed reasonable, the project manager is to be encouraged to reconsider the way AI should be developed, provided, and used. The list for evaluating deviations is prepared in accordance with Action Objective 3-1-1, referring to the standard deviation evaluation in the industry to which the Company belongs and this guideline, but the actual projects are selected and management accompanies the project personnel in order to make the list more precise and to establish the operation of the list. We are also devising ways to make the list more precise and to make its operation more established. The Committee on AI Governance will request project personnel to report the results of their reconsideration, and if there is any concern about the reasonableness of the contents of the report, the officer in charge of AI Governance will notify the officer in charge of the project to that effect and make adjustments accordingly.

Since the risks associated with AI systems and services vary greatly depending on the application, scope, and mode of use, and the person in charge of promoting the project is considered to know the nature and extent of the risks best, it may be possible for management

to attend the project meeting when it is clear that the potential risks are minor, and to conduct a simple deviation evaluation. Therefore, it is conceivable to operate without uniformly requiring strict evaluation of deviations, such as by using a simplified evaluation of deviations. However, at this point in time, the know-how for evaluating deviations and risks has not yet been sufficiently accumulated within the company, and therefore, we have decided to make a uniform deviation evaluation by the Committee on AI Governance a mandatory gate through which all projects should pass, and will see how the process progresses in the future.

Practical example iii: Divergence evaluation process with collaboration among various entities.

The process of deviation evaluation may need to be carried by more than one company. For example, when an AI provider offering services to others outsources the development of its AI system to an AI developer, rather than developing the AI system itself, it may be reasonable for both the AI developer and the AI provider to share the deviation evaluation process. In this case, it is important for the AI developer and the AI provider to share the methods and criteria for deviation evaluation as well as the flow expected from the development of the AI system to its operation. if the AI provider neglects the risks involved in providing services using the AI system, the AI developer will be placed in a difficult position This is important because if AI providers underestimate the risks associated with providing services using AI systems, AI developers will be placed in a difficult position.

Although we, who sometimes develop AI systems on consignment, have an agreement that any accident in the operation of the AI system is to be borne by the AI provider who provides the service, except in cases of circumstances attributable to us, we are still at risk of being involved in a dispute when an accident of this kind occurs. However, there is still a risk that we may be involved in a dispute when an accident of this kind occurs. Therefore, we cannot be indifferent to the operational methods of the AI systems we have delivered. In fact, we have had the experience of noticing operational risks in the final stages of a project, advising the AI provider to redesign the project in question, and then being forced to bear part of the cost of such redesign. Therefore, we established a deviation evaluation process based on a thorough understanding of the meaning of individual evaluation items, referring to the standard deviation evaluation in the industry to which we belong and these guidelines, and shared it with AI providers who do not develop their own products but only provide services to others. By utilizing the deviation evaluation process, which covers all items of concern, and by conducting deviation evaluations early, negotiations with customers have become smoother.

In some cases, as in the following practical example, it may be necessary to have a broad discussion in addition to the usual deviation evaluation process.

Example iv: Additional measures to be taken by small businesses in light of the risks of AI.

The firm is a small company whose primary business is the development of AI systems. The technical director receives progress reports on all projects, including reports on AI ethics, such as fairness, In socially sensitive areas, however, this may not be sufficient.

Therefore, in the case of an AI system project in such a sensitive area, we will discuss the project with our legal officers and others. In identifying sensitive areas, we refer to the thinking of leading companies that have already developed, provided, and used AI systems and services extensively. Practical journals are useful for gathering such information[25] . Such magazines often contain overview articles, and it is efficient and effective to use the overview articles as clues to access in-depth information on the Internet and other sources.

We know that some companies invite outside experts and specialists to exchange opinions on individual projects. As we expand our business, we would like to establish such a forum for the exchange of opinions.

[25] Reference examples of responses to sensitive areas include Satoshi Funayama, "Corporate Approaches to AI Ethics (1)," NBL No. 1170 (May 2020).

Practical example v: In addition to each entity, deviation evaluation in collaboration with external experts as necessary.

We are a large company with a mix of departments developing and operating AI systems and services. We already have an AI policy in place, and we evaluate all projects for deviations from that policy. If the project is in an area that has been handled in the past, it is sufficient for management to take action at an early stage of the project. We try to have them consult with us individually. And when such consultation is received, a cross-sectional meeting consisting of responsible persons from the development, operation, and legal departments is to be held to discuss the issue. The same applies when management discovers such projects during the normal deviation evaluation process.

We regularly invite outside experts and specialists to catch up on recent AI incidents and information on sensitive areas at an early stage. Therefore, for now, we can adequately respond to the situation by discussing it in cross-functional meetings based on the information and general advice obtained from the experts and specialists. On the other hand, as the applications and destinations of our AI systems and services are expanding, we believe that it may become necessary to seek opinions from outside experts and others regarding individual projects in the future.

Action Objective 3-1-1 [Ensure consistency with industry standard deviation assessment process]: Each entity, under the leadership of its management, is expected to confirm the existence of a standard deviation assessment process in the industry and incorporate such a process into its own process, if one exists.

[Points of practice].

Each entity is expected to work under the leadership of management to

- Proactively incorporate external best practices, such as industry-standard deviation evaluation processes and initiatives of other companies and organizations, without limiting our own knowledge and experience.

In addition to guidelines that can be used as a reference in each industry, it is also useful to check information from ministries and organizations that are relevant to your company, as some ministries and industry associations may have published guidelines on AI reliability assessment.

Examples include.

- Ministry of Economy, Trade and Industry "The State of AI Governance in Japan"[26]
- Ministry of Economy, Trade and Industry, Ministry of Health, Labour and Welfare, Fire and Disaster Management Agency "Guidelines for AI Reliability Assessment in the Plant Safety Field".[27]
- National Institute of Advanced Industrial Science and Technology (AIST) "Machine Learning Quality Management Guidelines".[28]
- Personal Data +α Study Group "Checklist on Voluntary Initiatives" in Profiling
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3".[29]
- NIST, "AI Risk Management Framework Playbook."[30]

[26] Ministry of Economy, Trade and Industry "AI Governance in Japan ver1.1" (July 2021)
[27] Ministry of Economy, Trade and Industry, Ministry of Health, Labour and Welfare, and Fire and Disaster Management Agency, "Guidelines for AI Reliability Assessment in the Plant Safety Field, Version 2" (March 2021),".
https://www.meti.go.jp/press/2020/03/20210330002/20210330002-2.pdf
[28] National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023),.
https://www.digiarc.aist.go.jp/publication/aiqm/AIQuality-requirements-rev4.1.0.0112-signed.pdf
[29] Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, "Corporate Privacy Governance Guidebook in the DX Era ver. 1.3" (April 2023),
https://www.soumu.go.jp/mAIn_content/000877678.pdf
[30] NIST, "AI Risk Management Framework Playbook" (January 2023), https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

- EU, Assessment List for Reliable AI[31]
- IMDA, "AI Verify An AI Governance Testing Framework and Toolkit."[32]
- Financial Services Agency "Principles for Model Risk Management[33]

[Practical example].
[Practical example i: Incorporating the deviation evaluation process of other companies and organizations' guidelines

Since diverse perspectives are essential in practicing AI governance and sharing recognition with other companies is also necessary, we should refer to the efforts of other companies and organizations, rather than thinking only of our own. In light of this belief, the management of Minebea has instructed its governance staff to investigate the efforts of other companies in establishing a deviation evaluation process.

Since our main business is the development of AI systems for industrial applications, we conducted a survey focusing on industrial applications. In the course of our research, we found, for example, that the Ministry of Economy, Trade and Industry, the Ministry of Health, Labor and Welfare, and the Fire and Disaster Management Agency have published "AI Reliability Evaluation Guidelines for Plant Safety Field," "Implementation Details Recording Format" for implementing the guidelines[34] , and "Practical Examples of Reliability Evaluation"[35] , which contains examples of descriptions. We also found that the "AI Product Quality Assurance Guidelines[36] " published by the AI Product Quality Assurance Consortium includes examples of Voice User Interface, industrial process, automatic operation, and OCR. We also found that the National Institute of Advanced Industrial Science and Technology (AIST) has published "Machine Learning Quality Management Guidelines" and has created a reference guide as specific application examples for actual applications by industrial use. In addition, the guide includes the process of quality management according to the guidelines and the form of the Machine Learning Quality Management Guideline Assessment Sheet suitable for planning and recording the process, as well as the instructions for its use.[37] Our current deviation assessment process reflects some of these specific efforts.

[Practical example ii: Incorporating considerations for handling personal information into the deviation evaluation process].

We develop, provide, and use AI systems and services based on data obtained from non-business users, and we recognize that the practice of AI governance, especially in ensuring privacy, requires not only consideration for the construction of AI models and outputs, but also for the handling of input data for AI models. In particular, we recognize the need to consider not only the construction of AI models and their outputs, but also the handling of input data to AI models. Although the Company has abundant experience in handling personal information, even so, we believe it is important to actively look at efforts outside the Company. Therefore, management has instructed the privacy officer to investigate external efforts in constructing the deviation evaluation process.

Regarding the construction of AI models and consideration of outputs, we found, for example, the "Checklist for Voluntary Initiatives" in profiling presented by the Personal Data +α

[31] EU, "Assessment List for Trustworthy Artificial Intelligence (ALTAI)" (June 2020), https://digital-strategy.ec.europa .eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
[32] The Infocomm Media Development Authority, "AI Verify - An AI Governance Testing Framework and Toolkit" (2022). (May), https://aiverifyfoundation.sg/
[33] FSA, "Principles for Model Risk Management" (November 2021), https://www.fsa.go.jp/common/law/ginkou/pdf_02.pdf
[34] Ministry of Economy, Trade and Industry, Ministry of Health, Labour and Welfare, and Ministry of Internal Affairs and Communications, "Reliability Assessment Implementation Record Format," https://www.meti.go.jp/press/2020/03/20210330002/20210330002-3.pdf
[35] Ministry of Economy, Trade and Industry, Ministry of Health, Labour and Welfare, and Ministry of Internal Affairs and Communications, "Overview of Reliability Assessment Practical Examples (7 examples)," https://www.meti.go.jp/press/2020/03/20210330002/20210330002-4.pdf
[36] AI Product Quality Assurance Consortium, "AI Product Quality Assurance Guidelines Version 2023.06" (June 2023), https://www.qa4AI.jp/QA4AI.Guideline.202306.pdf
[37] National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Reference Guide" (July 2022), https://www.digiarc.aist.go.jp/publication/aiqm/referenceguide.html

Study Group[38] . As for the "Corporate Privacy Governance Guidebook in the DX Era ver1.3"[39] , we found it helpful as it also describes AI in terms of both inputs and outputs. Our current deviation assessment process reflects some of these specific efforts.

Action Objective 3-1-2 [Provide AI users and off-business users with sufficient information on possible deviations/measures to deal with such deviations]: Under the leadership of management, each entity is expected to provide stakeholders with sufficient information on the fact and countermeasures against possible deviations from the AI systems/services it provides, as well as to clarify the contact point for inquiries.

[Point of practice.
Each entity is expected to work under the leadership of management to

- If there is a possible gap between AI systems/services and "AI Governance Goals," provide stakeholders with information on the fact and measures to address the gap, and communicate with them through responses to their inquiries, etc.
- To enhance the effectiveness of information provision, we will also contribute to improving the literacy of AI users and non-professional users through various information dissemination, in cooperation with AI developers and industry associations.
- Consider the degree of information to be provided in accordance with the nature and probability of the risk created by the discrepancy.

A specific example of providing information tailored to stakeholder literacy is the choice of terminology.

- If there is a wide range of literacy, mention the basic structure of the AI system/service and explain it in plain language so that all stakeholders can understand it.
- For stakeholders with high literacy, explain in a crisp manner with technical terms.

In addition, the following are examples of communication through inquiry reception.

- Clearly state the contact information as a prerequisite.
- Clearly state that AI is used in the system in an easy-to-understand location, such as on the website.

[Practical example].
[Practical example i: Provision of information with reference to the "Corporate Privacy Governance Guidebook in the DX Era ver1.3"].
The Company operates AI systems/services and provides AI services to an unspecified number of off-business users. Given that there is expected to be a wide range in the AI literacy of service providers, we provide information related to risks, such as appropriate risk management and measures to minimize risks in operating AI systems and services, and strict safety control of information, in a clear and easy-to-understand manner so that even non-business users unfamiliar with AI can understand it. Information is provided in an easy-to-understand format so that even non-business users unfamiliar with AI can understand it, and the contact point for inquiries is clarified. In addition to this information, as mentioned above, we expect that there is a wide range of literacy regarding AI among service providers. The Company clearly indicates the advantages and disadvantages of using AI, as well as indicating the advantages and disadvantages of using AI in an easy-to-understand manner. For non-business users who do not prefer AI, alternative services are also indicated. Since personal information may be handled in some cases, the company not only complies with the Personal Information Protection Law and the guidelines of the Personal Information Protection Commission, but also establishes continuous communication with non-business users, referring to the "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (in Japanese). The

[38] Personal Data + Alpha Study Group "Final Recommendations on Profiling" p.10-21 (April 2022)
[39] Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, "Corporate Privacy Governance Guidebook in the DX Era ver. 1.3" (April 2023), https://www.soumu.go.jp/mAIn_content/000877678.pdf

company has also established a "Privacy Governance Guidebook for Companies in the DX Era ver. 1.3".

[Practical example ii: Providing information tailored to the AI literacy of the recipient (when the recipient's AI literacy is high)

As in Practice i, the Company operates AI systems/services and provides AI services to external parties, but differs from Practice i in that the Company provides AI services to companies that use its services for business purposes. Since the recipients of our services have a relatively high level of AI literacy, we explain in a well-defined manner, including technical terms, the possibility that a certain level of deviation may occur in the AI systems/services we provide and the measures to deal with such deviation, as well as clarify the contact point for inquiries.

In the future, we may provide AI services using AI systems for non-business users, and we would like to provide sufficient information depending on the AI service providers' literacy in AI.

[Practice iii: Providing information tailored to AI literacy (if there is a range of AI literacy in the recipient)

The Company is taking the same measures as in Practice i. However, the Company believes that enabling AI users and non-business users to select AI services using AI systems at their own discretion has added value in itself, and is devising ways to provide information to differentiate itself from other companies. In addition, the company is devising ways to receive feedback not only on AI systems and services but also on the way information is provided.

Example iv: Collaboration with AI developers, etc.

We are taking the same measures as in Practice i, but we make it clear in the contract that the AI developers will provide us with the information necessary to respond to inquiries from AI users and off-business users. The AI developers respond to them promptly, as the "feedback" from AI users and off-site users is also valuable information for them.

Column 2 : Data providers to provide sufficient information on discrepancy assessment to AI developers

Data providers are expected to provide information on data sets, including data collection sources, collection policies, collection criteria, annotation assignment criteria, and usage restrictions, to enable AI developers and AI providers to properly evaluate deviations, and AI developers are expected to obtain data sets from data providers that provide sufficient information The AI developer is expected to obtain datasets from data providers that provide sufficient information. In the case of LLMs for generative AI, there may be limitations in providing information on datasets, so as much information as possible should be provided by AI service providers, and this fact should be shared with relevant stakeholders.

[Notes]

AI systems and services are highly dependent on the underlying data for their fairness and other qualities. Therefore, it is a prerequisite for AI developers and providers to receive sufficient information on the source data from the data providers in order to properly evaluate deviations.

The following information on the data set is applicable.

- Data Collection Policy: Approach to data collection, etc.
- Source of data collection: Original data source/provider, scope of data collection, etc.
- Data collection policy: objects and items to be collected, methods of collection, period of time during which the data was collected, etc.
- Data collection criteria: conditions of data collected, cleansing methods, data bias, etc.
- Data annotation assignment criteria: annotation rules for image/audio/text, etc.
- Restrictions on use of data: restrictions derived from other rights, etc.
- Purpose of data use: In particular, in the case of data containing personal information, the specific purpose presented to the individual concerned, etc.

[Practical example].

As a data provider providing data to AI developers and AI providers, we provide information on data sets, including data collection sources, collection policies, collection criteria, annotation assignment criteria, and usage restrictions, in order to enable companies developing AI systems to conduct deviation assessment appropriately. Even when a data set that is not sufficiently organized is provided, basic information such as the collection source of data necessary for deviation evaluation is sufficiently provided.

Action Objective 3-2 [Improve human resource literacy for AI management systems]:
Each entity is expected to strategically improve AI literacy under the leadership of management, including consideration of the use of outside educational materials for appropriate management of AI management systems. For example, the officers, management team, and personnel responsible for the legal and ethical aspects of AI systems and services should be trained to improve their general literacy in AI ethics and AI reliability, and those in charge of projects to develop, provide, and use AI systems and services should be trained not only in AI ethics but also in generative AI Training on AI technology, including AI technology, as well as AI ethics, could be provided to all personnel, and education on the position and importance of AI management systems could be provided to all personnel.

[Points of practice].

Each entity is expected to work under the leadership of management to

- To improve AI literacy by using training and educational materials appropriate for the position and responsibilities, including those provided by outside instructors.
- In doing so, utilize training and educational materials appropriate to the role to be fulfilled by each person.
- Ensure that all employees take courses on AI ethics, which are particularly important.

- In light of recent trends in generative AI, provide training on generative AI technology and the reliability of output results, etc.
- Consideration to enable employees to acquire such expertise when evaluation by persons with relevant expertise independent of the design and operation of the AI management system under Action Objective 5-1 is conducted in-house.

Note that the required AI literacy will change as AI technology advances, so the mismatch between human resource development and the speed of technological change should be kept in mind.

[Practical example].
Example of practice i: Education using external teaching materials, etc.

Since MHI is a small company and the number of trainees is small, the company decided to use outside educational materials instead of preparing its own training program to improve AI literacy. Various educational programs are available, both domestically and internationally, including online courses and textbooks provided by Coursera, a for-profit educational technology organization in the U.S., the Japan Deep Learning Association (JDLA), and others, as well as "Manavi DX"[40] and "Manavi DX Quest"[41] from the Ministry of Economy, Trade and Industry.

MHI utilizes a program based on the JDLA's certification syllabus to measure the achievement of its trainees; the JDLA's G certification exam covers a wide range of topics, from the fundamentals of AI technology to AI ethics. In a survey of successful trainees of the JDLA's G2023#3 (conducted on July 7, 2023), the largest number of successful trainees (30%) responded that their study time was 15-30 hours,[42] confirming that this is not an undue burden on the trainees.

In addition, the Company believes that improving the digital literacy of its employees is essential for utilizing digital technology, including AI, and recommends that all employees throughout the company obtain the "IT Passport" .[43]

We believe that the program has had the desired effect that we have implemented so far. For example, people who had only heard bits and pieces in the news about incidents of AI systems and services have mastered the rudiments and ethical aspects of AI technology, and are now thinking about AI risks with a sense of ownership.

Example ii: Education using in-house teaching materials

We are a large company that develops, provides, and uses AI systems and services as one of the pillars of our business. we know that there are external educational materials on AI technology and ethics, but due to the large number of AI systems and services we provide and the benefits/risks to society, we do not use generic external educational materials, but our own AI We are using our own teaching materials, which are enriched with case studies for the intended use of our systems *and* services. In addition, for practical in-house education using AI, the company also uses case study materials with data in "Manavi DX Quest" by the Ministry of Economy, Trade and Industry .[44]

When the training program on AI was first created, a section on AI ethics was included at the end of a lecture on AI technology. However, a committee of outside experts pointed out that AI ethics was of growing interest to management, and an e-learning course on AI ethics alone was created and made available to all employees. This e-learning includes lectures. The e-learning includes a lecture and a confirmation test, and is designed to be completed in about one hour

[40] Ministry of Economy, Trade and Industry, "Manabi DX Home Page," https://manabi-dx.ipa.go.jp/
[41] Ministry of Economy, Trade and Industry, "Manabi DX Quest Home Page," https://dxq.manabi-dx.ipa.go.jp/
[42] Japan Association for Deep Learning, "Interview with G-Certificate Successful Candidates" https://www.jdla.org/certificate/general/start/
[43] Information-technology Promotion Agency, Japan "IT Passport Examination Home Page", https://www3.jitec.ipa.go.jp/JitesCbt/index.html
(Questions related to generative AI are scheduled to be asked from FY2024.)
[44] Ministry of Economy, Trade and Industry, "Manabi DX Quest 'Provision of Case Study Materials with Data'," https://www.meti.go.jp/policy/it_policy/jinzai/manabi-dx-quest.html

even by those who are not familiar with AI ethics. By relating the training to the applications of its AI systems and services, the company believes that it has achieved a high learning effect even in a short period of time.

Example iii: Education on Generative AI

Most recently, we believe that it is also necessary to develop human resources to deal with generative AI. Referring to the Ministry of Economy, Trade and Industry's "Concept of Human Resources and Skills Required to Promote DX in the Generative AI Era"[45] and "Digital Skill Standards"[46] , we recommend our employees to take e-learning courses related to generative AI posted on "Manavi DX" and are also considering using JDLA's The company is also considering the use of the JDLA Generative AI Test, a certification test to test the ability and knowledge to properly use generative AI.

Action Objective 3-3 [Strengthen AI management through cooperation among various entities and departments]:
Except for cases in which each entity conducts everything from preparation of datasets for learning to development, provision, and use of AI systems and services by its own division, each entity is expected, under the leadership of its management, to clearly identify operational issues of AI systems and services that cannot be fully implemented by itself or its own division alone, while paying attention to trade secrets, etc., and to share information necessary to solve such issues to the extent possible and reasonable under the conditions of ensuring fair competition. In such cases, it is expected that necessary information should be shared to the extent possible and reasonable under the condition of ensuring fair competition. In doing so, it is expected that each entity should agree in advance on the scope of information disclosure and consider concluding a nondisclosure agreement, etc., to ensure smooth exchange of necessary information.

[Points of practice].

Each entity is expected to work under the leadership of management to

- Identification of operational issues of AI systems and services that cannot be resolved by each entity alone and the information needed to resolve them
- Sharing among entities to the extent possible and reasonable, while paying attention to intellectual property rights, privacy, etc.
- The above is based on the assumption that fair competition is ensured by various laws and regulations, AI policies of each entity, trade secrets, limited provision data, etc.

In addition to the relevant laws and regulations such as the Unfair Competition Prevention Law and the Personal Information Protection Law, contracts between entities are also relevant, so legal and risk compliance staff should be consulted (for details, see "Appendix 6. (For details, please refer to "Appendix 6.)

In addition, when each entity spans multiple countries, the status of the international community's consideration of appropriate AI governance to ensure free cross-border transfer of data (Data Free Flow with Trust, hereinafter referred to as "DFFT") and interoperability based on it (the two aspects of "standard" and "interoperability among frameworks")), and implement risk management and AI governance appropriate for each stage of development, provision, and use, as well as clarification of the risk chain, including data distribution.

[Practical example].
[Practical example i: Careful information sharing for customers who are not familiar with AI.

[45] Ministry of Economy, Trade and Industry, "Concept of Human Resources and Skills Necessary to Promote DX in the Generative AI Era" (August 2023), https://www.meti.go.jp/press/2023/08/20230807001/20230807001-b-1.pdf
[46] Ministry of Economy, Trade and Industry, Information-technology Promotion Agency, Japan "Digital Skill Standards ver.1.1" (August 2023), https://www.ipa.go.jp/jinzai/skill-standard/dss/index.html

The Company delivers the AI systems it has developed to its customers, who in turn operate the AI services. The accuracy of this AI system/service may deteriorate due to changes in the operating environment, and in some cases, this may lead to damage to equipment or other damage. Therefore, we request our customers to monitor the output of the AI system and also inform them how to judge quality deterioration.

For customers who are not familiar with AI, simply requesting monitoring and other services will not work; it is necessary to take time to explain and convince them of the reasons why maintenance of AI systems and services is necessary, the causes (e.g., changes in the distribution of training data and input data during operation), and the trends in output changes caused by such causes. It is necessary to take time to explain and convince them of the reasons and causes (e.g., changes in the distribution of training data and input data during operation). Although providing standard information may be sufficient in some cases, even if the AI system/service developer thinks so, it is important to actively encourage the supplier to ask questions and to make their understanding as consistent as possible. If necessary, it is also important to conclude a maintenance service agreement and establish a system to proactively accept questions even after delivery. When an AI system/service is re-learned, it is important to carefully explain how the output has changed as a result of the re-learning. For example, as a point to keep in mind in relearning, it is important to explain the "model failure," a degradation problem that occurs when data obtained by AI as output is used as input for relearning (a phenomenon in which AI repeatedly learns its own errors, and as these errors gradually accumulate, the performance of AI systems and services gradually deteriorates), and so on. (AI repeatedly learns its own errors, and the performance of AI systems and services gradually deteriorates as these errors accumulate).

Example ii: Execution of nondisclosure agreements for smooth information sharing.

To ensure smooth sharing of the above information, MHI has agreed in advance on the scope of information disclosure between AI developers and AI providers, and has also concluded a nondisclosure agreement.

[Practical example iii: Thorough information sharing through additional verbal explanations

The AI system we are developing is trained by a specific data set, and may produce undesirable output results when applied to a target not included in the data set. Therefore, AI providers who intend to offer such AI systems to AI users are not only informed about the data used for learning, etc., the outline of models used, and performance such as accuracy, but also about situations and targets for which the AI systems should not be used. In order to ensure thorough provision of information, the information is not only conveyed in paper or electronic written form, but is also explained orally, reserving a separate time for such explanation and having the user sign that he/she has received such an explanation.

Practical example iv: Information sharing across multiple countries

The Company is an AI developer and provider headquartered in Japan. As a global provider of AI systems and services, we believe that more careful coordination is essential for risk management when AI users and non-business users are registered overseas. In particular, it is important to take into consideration the social differences in culture, climate, and acceptability regarding AI in each country.

In addition, in order to comply with legal regulations that differ from country to country, we research laws equivalent to personal information protection laws and regulations on data security in the countries where AI users and non-business users are located, and build security measures, etc. based on these laws and regulations.

In addition, we are gathering information on the international debate on DFFT and the various frameworks for data distribution that may affect our business, using experts in the field.

Column 3 Example of consideration for social differences in culture, climate, acceptance of AI, etc. in each country
social differences in culture, climate, acceptance of AI, etc. in each country.

An example of an actual response is the Hub & Spoke model adopted by Microsoft Corporation (the Spoke portion is handled by appointing an AI Champ from the country/region where the service is provided, while incorporating the perspective of that country/region).[47,48]

Another example of the company's multi-stakeholder engagement is the Global Perspectives Responsible AI Fellowship, which was launched in partnership with the Stimson Center's Strategic Foresight Hub. The purpose of the fellowship is stated to be to invite stakeholders from the Global South countries into various discussions on AI. [49,50]

Action Objective 3-3-1 [Understanding the current situation through information sharing among each entity]:
Each entity is expected to understand the current status of relevant information sharing among the entities and to update its understanding in a timely manner, paying attention to trade secrets, except in cases where the entity, under the leadership of management, conducts everything from preparation of datasets to be used for learning, etc., to use of AI systems and services on its own.

[Points of practice].
Each entity is expected to work under the leadership of management to

- Sharing of information on sources of data acquisition/data quantity and quality, distribution, and overview of each category of data used in the development of the AI system.
- When sharing information, refer to the National Institute of Advanced Industrial Science and Technology's "Machine Learning Quality Management Guidelines" and other efforts to standardize information sharing.

In doing so, each entity is expected to correctly understand the following Therefore, it is expected to standardize the shared information in order to facilitate information sharing among the various entities and promote social implementation of AI technologies.

- Development of AI systems is envisioned for situations of provision and use.
- AI systems are provided with a proper understanding of the constraints under which the AI system was developed and how it will be used as a service
- Use of AI services is conducted with an understanding of the intended use of the AI provider and within the scope of such use.

Methods of sharing/obtaining information include

- Confirmation of guidelines established by relevant ministries, agencies, industry associations, etc.
- Membership in AI ethics and quality organizations
- Refer to precedents, including overseas
  - References to professional organization reports
  - Participation in seminars, etc.

[47] Microsoft, "More on Microsoft's Responsible AI Program : Governance as the Foundation for Compliance" (February 2023), in https://news.microsoft.com/ja-jp/2021/02/02/210202-microsoft-responsible-ai-program/
[48] Microsoft, "The building blocks of Microsoft's responsible AI program : Governance as a foundation for compliance " (January 2021), https://blogs.microsoft.com/on-the-issues/2021/01/19/microsoft-responsible-ai-program/
[49] Microsoft "Promoting AI Governance in Japan : AI Governance within Microsoft" (October 2023), https://news.microsoft.com/ja-jp/2023/10/06/231006-about-the-potential-of-ai-in-japan/
[50] Microsoft, "Advancing AI governance in Japan : Governing AI within Microsoft" (October 2023), https://blogs.microsoft.com/ on-the-issues/2023/10/05/responsible-ai-governance-japan/

- Interviews with experts, etc.

[Practical example].
Practical example i: Efforts to standardize information sharing among various entities

In deciding how its own information should be provided, the Company, under the leadership of its management, has decided to understand the current status of information sharing among the various entities and to periodically update its understanding, while being mindful of trade secrets.

As we gathered information, we also learned that various efforts are being made to standardize information sharing among various entities. For example, the National Institute of Advanced Industrial Science and Technology (AIST) has published the "Machine Learning Quality Management Guidelines," one of the objectives of which is to establish a social consensus standard for the quality of systems using machine learning. It was also learned that the Ministry of Economy, Trade and Industry, the Ministry of Health, Labor and Welfare, and the Fire Defense Agency have created a reliability assessment implementation record format for the safety field. We also understood that model cards have been proposed based on the recognition that it is important to display the performance of AI models as well, just as the labeling of food ingredients and other information contributes to responsible decision-making by people. [51]

At this point, there is no standard documentation procedure for sharing the performance and quality of learned machine learning models, etc. among the various entities, but in developing an internal system, we intend to refer to various initiatives rather than developing our own standards from scratch.

[Practical example ii: Understanding the current status of information sharing among various entities through organizations on AI ethics and quality.

We are a member of an AI ethics and quality organization, and actively exchange opinions with other companies we belong to on the appropriate way to provide information on the performance of AI systems and services, etc. It is important to provide AI users and non-professional users with sufficient information on AI systems and services. However, it is not appropriate to think that it is sufficient to unilaterally provide information that is difficult for non-specialists to understand or that is voluminous and detailed, since they are not familiar with the nature and limitations of AI systems and services. In order to consider the appropriate way to provide information, we consider it important to be in contact with many stakeholders not only through direct experience of the company itself, but also indirectly through exchanges of opinions with other companies.

Information that may need to be communicated from the AI developer to the AI provider includes, for example, information about the data used in the development of the AI system. This information could include, for example, the source from which the data was obtained (sometimes called open data), the quantity and distribution of the data, and an overview of each category included in this data. It is also important to provide an overview of the algorithms selected (or not) during development and the models generated, in particular, under what conditions they were tested and what level of accuracy was achieved as a result.

While these perspectives are not new to companies with extensive experience in developing and providing AI systems and services, we believe that "how to communicate" is critical. It is a question of what content to explain and in what depth. Understanding the current state of information sharing among various entities is important for the overall design of AI governance, and therein lies the significance of participating in organizations on AI ethics and quality.

Practical example iii: Cooperation between entities in the case of multiple countries

We are a company that develops and provides AI. Since we often collaborate with other entities, we regard information sharing as highly important and are actively investigating trends in advanced cases, including those overseas.

[51] Google, "Vertex AI," https://cloud.google.com/vertex-ai

First, we refer to reports from universities and specialized institutions. In addition, they refer to case studies of initiatives on the websites of advanced companies introduced in these reports.

In addition, information on social networking sites seems to have increased in importance in recent years: in addition to looking at posts on social networking sites and other online platforms, the company also refers to seminar announcements and other information, and actively encourages employees to participate in those that are closely related to the company.

In addition, the company regularly invites consultants and other experts in AI who are knowledgeable about the latest trends and case studies to its offices to receive advice on how to incorporate them into its strategies and what actions to take.

Action Objective 3-3-2 [Encouraging daily collection of information and exchange of opinions for environmental and risk analysis]:
Each entity is expected to routinely collect information on rule development, best practices, and incidents related to the development and operation of AI systems and services under the leadership of management, as well as to encourage the exchange of opinions internally and externally.

[Points of practice].

Each entity is expected to work under the leadership of management to

- Daily collection of information on rule development, best practices, incidents, etc.
- Even if an AI management team is established within the company, hold discussions and study sessions with other departments within the company, and get involved in group activities in which other companies also participate.

[Practical example].

Example i: Encouraging management-led discussions.

With regard to AI ethics, the guidelines are being established, but since there is no right answer as to how the guidelines should be respected, and since other companies are also engaged in similar activities, management encourages personnel in each department to gather information and exchange opinions on the appropriate development, provision, and use of AI, and instructs them to share this information in internal discussions and study sessions across departments. They are instructed to share this information in internal discussions and study groups across the company.

By continuing these activities, we have been able to identify major trends, although there is no definitive solution, and the results of these activities are reflected in the environmental and risk analyses that are conducted in a timely manner.

Example ii: Encouraging discussion in small businesses

We are a small company developing AI systems. Since there is an opinion within the company that growth should be more important than respect for AI ethics, we decided to start with internal discussions and study sessions on AI ethics in the legal and technical departments. Since the definition and usage of the term may differ from department to department, a facilitator was appointed to facilitate the discussion, which proceeded smoothly, and it became clear that the engineers, who said they should focus on growth, had already been exposed to papers dealing with fairness, etc., and there was no significant difference in their perception of AI ethics. Since engineers began to show interest in realizing respect for AI ethics through technology, the development process is changing to one consistent with AI ethics. In the future, we would like to promote the exchange of opinions with outside parties.

Action Objective 3-4 [Reduce the incident-related burden on AI users and off-the-job users through prevention and early response]:

Each entity is expected to reduce incident-related burdens on AI users and off-business users through prevention and early response to incidents under the leadership of management.

[Points of practice].
Each entity is expected to work under the leadership of management to

- Prevention and early response to incidents such as system failures, information leaks, and claims
- Establish a system to prevent incidents throughout the lifecycle or respond to incidents as soon as possible when they occur.

The following points should be considered when establishing a system to prevent incidents or to respond to incidents as soon as possible.

- Study preventive and preparedness measures by accumulating past cases and utilizing information collected in Action Objective 3-3.
- Distribution of responsibility among relevant entities (to those who can reduce risk)
- Early response to economic losses through the use of insurance for uses that have a certain probability of economic loss

[Practical example].
Practical example i: Clarification of where responsibility lies.
The development, provision, and use of AI systems and services often involve entities and individuals in various positions, such as AI developers, AI providers, AI users, and off-business users, etc. Furthermore, the so-called black box nature of AI makes it easy for responsibility to become unclear. To prevent incidents, it is important to distribute responsibility to those who can mitigate risks. Therefore, we have established a system that enables early response to incidents by clarifying who is responsible in the event of an incident and by granting a certain level of authority to the responsible person. It is also important to enhance the ability to respond to incidents as early as possible by preparing in advance for the occurrence of incidents.

[Practice ii: Use of insurance for incidents and ongoing R&D].
While the Company is basing the implementation of Practice i, the Company is considering the use of insurance for some applications. For applications where a certain degree of uncertainty regarding the operation of AI systems and services is inevitable and a certain amount of economic loss occurs in rare cases, despite the significant benefits to society as a whole, we believe it is important to reduce the burden on AI users and off-business users by using insurance to respond early to possible economic losses from the incident. We believe it is important to reduce the burden on AI users and off-business users by using insurance to respond to possible financial losses from incidents as early as possible. Of course, we recognize the importance of reducing the uncertainty of AI systems and services in order to continuously increase the trust of AI users and out-of-business users, and are continuing research and development to this end.

Action Objective 3-4-1 [Distribution of the burden of responding to uncertainty among the various entities]:
Each entity is expected to clarify where the responsibility for dealing with uncertainties in AI systems and services lies so that risks can be minimized across the board under the leadership of management.

[Points of practice].
Each entity is expected to work under the leadership of management to

- To begin with, we recognize the premise that it is difficult to completely eliminate uncertainty in AI systems and services, although it is technically possible to deal with some of it [52]
- Then, clarify where the responsibility lies among each entity to the extent possible and reasonable.

Contracts and other agreements may be effective in clarifying where responsibility lies among the various entities.

There is a debate as to which of the various entities needs to guarantee the quality of AI systems/services, and the situation differs for each AI system/service; however, "Appendix 6. Key Considerations When Referring to the "Guidelines for Agreements on AI and Data Use"" in these Guidelines is also helpful.

In addition, if the value chain/risk chain from AI development to the provision of AI-based services is expected to span multiple countries, consideration of appropriate AI governance for cross-border transfer of data, data localization, etc. should also be noted.

[Practical example].

Practical example i: Responding to uncertainty through information linkages to other entities.

As an AI system developer, we believe that having AI used by relevant stakeholders will contribute to improving public trust in AI technology. In gathering information, we found that some AI providers consider AI systems/services to be an extension of conventional software and believe that AI developers should bear all responsibility for the quality of AI systems/services. On the other hand, it was found that AI providers themselves may be able to determine the timing of relearning by understanding the expectations of AI systems/services and explaining them carefully until the AI providers themselves feel comfortable with them. We also understood that the concept of the importance of "quality assurance engineers, teams, and organizations working together with development and sales to deepen customers' understanding of AI systems" as described in the "AI System Quality Assurance Guidelines" is gradually spreading. However, the idea that AI developers should assure quality still persists, and we would like to continue to conduct surveys on the burden of dealing with uncertainty on a regular basis, hoping that the positive impact of activities such as the "AI System Quality Assurance Guideline" will spread.

Practical example ii: Prepare an explanation in case of a claim of responsibility.

The Company is an AI provider offering AI services using AI systems developed by other companies; it concludes contracts with AI developers using model contracts based on the "Contract Guidelines for the Use of AI and Data" at[53] . According to this, AI developers of AI systems and services (learned models) are required to perform their work with a certain level of care and attention, while not guaranteeing the completion of the work or the performance and quality of the results. We are aware that we are only operating AI systems/services developed by other companies, and if any inappropriate cases occur in relation to the operation of AI systems/services, or if we are asked for explanations by users outside our business in other situations, we, as an AI provider, must be aware of what kind of We did not seriously consider the importance of fulfilling accountability.

However, regardless of the ultimate legal responsibility, as we are the direct provider of services to AI users, we cannot be exempted from any and all responsibility to respond to such requests, at least in the first instance, when AI users ask for explanations about AI systems and services we are operating, and we have changed our policy on what AI providers can do to reduce the risk, in cooperation with AI developers. After realizing that we cannot avoid such responsibility, and that we will be exposed to reputational risk if we fail to provide sufficient

[52] There are approaches that aim to reduce uncertainty through actions taken during the development of AI systems, such as preparing appropriate data sets, selecting appropriate models, and conducting validation and testing prior to the start of AI system use.

[53] Ministry of Economy, Trade and Industry, "Contract Guidelines for the Use of AI and Data" (June 2018) , https://www.meti.go.jp/policy/mono_info_service/connected_industries/sharing_and_utilization /20180615001-1.pdf

explanation, we have changed our policy to do what AI providers can do to reduce the risk and explain this as necessary, with the cooperation of AI developers.

Practice iii: Dealing with Uncertainty Regarding Data Handling

MHI was planning to outsource the development of an AI system using data held by MHI to another company, but MHI lacked expertise in data handling and wanted to entrust other companies with not only the pre-processing of data, such as cleansing, but also ensuring data quality. MHI mistakenly believed that if it gathered the data it currently possessed and provided it to the AI developer of the AI system, the AI developer, a professional in handling data, would perform the necessary processing on the data and develop the AI system/service that MHI desired.

However, as we collected information before commissioning development, we found that there is a "Practical Guidebook on Data Provision for AI/Data Science Human Resource Development" ([54]), which is also a reference for data provision among general entities, and it includes the concept that "only the commissioner can control the quality of commissioned data before provision. It also states that, under certain assumptions, "the profit from the use of the results also belongs only to the consignor...based on the concept of risk and reward liability...the consignor is in principle responsible for any damage caused by the use or implementation of the results created. The report summarized the points for AI developers to keep in mind, such as that there are cases in which "the profits from the use of the results are attributed to the consignor.

Currently, we understand that the content of data required for the development of an AI system will be determined by the nature of the AI system/service we plan to develop, and that there are limits to what can be handled on the part of the AI developer. We would like to pause here to reconsider the burden of dealing with uncertainties among various entities, given that even the data provision stage is an important part of the lifecycle of development, provision, and use of AI systems and services.

Practical Example iv: Dealing with Uncertainty in Generative AI

The Company is a company that develops and provides AI systems and services using generative AI to AI users, including those in other countries.

First, we are mindful of the increased potential for problems to arise in terms of rights relationships, including copyrights, in generative AI, and we place emphasis on concluding clear and fair agreements regarding copyrights and other rights related to training data and generative models. We recognize that the multinational nature of the data used in the development process may give rise to rights relationships based on different legal frameworks. For this reason, the Company is taking stock of the relevant laws, regulations, and risks by consulting with experts. In addition, we clarify the scope of responsibility with AI users. In doing so, we document the review process and ensure transparency so that we can move smoothly to resolve any legal issues that may arise.

Practice v: Dealing with Uncertainty When Spanning Multiple Countries

MHI is also considering AI governance for cross-border data transfers and data localization to address the challenges posed when the AI value chain/risk chain spans multiple countries. In doing so, the Company takes necessary actions depending on the nature of the AI services it provides and the scale of the possible risks it may pose, after checking the laws and regulations of each country involved, while seeking advice from experts.

We have also begun to consider different data storage methods as a risk hedge to allow us to respond flexibly to changes in international regulations. Specifically, we are considering deploying data centers in different regions to be able to respond to legal requirements for data handling in specific countries, utilizing the cloud to ensure flexibility to apply to legal changes in different countries, and also considering decentralized processing of data to smoothly

[54] Ministry of Economy, Trade and Industry, "Practical Guidebook on Data Provision for AI and Data Science Human Resource Development" (March 2021)

respond to different legal environments in different countries by separating data movement and processing. Decentralized processing of data is also being considered for this purpose.

Action Objective 3-4-2 [Advance review of response to incidents]:
Under the leadership of management, each entity is expected to consider deciding on a response policy and formulating a plan to promptly explain to AI users and non-business users, identify the scope of impact and damage, organize legal relationships, and consider damage relief measures, measures to prevent damage from spreading, and measures to prevent recurrence, etc., when an AI incident occurs, as well as to conduct practical rehearsals on the response policy or plan as appropriate. In addition, practical rehearsals are expected to be conducted on the response policy or plan, as appropriate.

[Points of practice].
Each entity is expected to work under the leadership of management to

- Develop response policies and plans in the event of an AI incident
- Conduct practical preliminary exercises on the above, as appropriate

The following systems are expected to be in place in advance in case of an AI incident.

- Establishment of a contact desk
- Assignment of officers in charge of response
- Assignment of roles to individual responders
- Response approach/process
- System for communicating with risk management divisions and other relevant internal parties
- System for contacting outside parties and experts, such as legal counsel
- Process for notifying stakeholders, etc.

If the impact of an AI incident on the business of an AI system or service is significant, consideration should be given to including the AI incident as a critical element in actually triggering the business continuity plan (BCP).

[Practical example].
Practical example i: Establishment of a system to prepare for incidents in small businesses].
As a small and medium-sized company providing AI systems and services, we recognize that while it is of course important to minimize as much as possible the likelihood of an AI incident occurring, it is difficult to reduce the likelihood of an AI incident to zero, and therefore it is important to develop and implement a plan to minimize damage when an AI incident occurs. Therefore, it is recognized that it is important to formulate and activate a plan to minimize damage in the event of an AI incident.
Specifically, in preparation for the event of an AI incident, the company has established a contact point, assigned an executive in charge of response, and established a system for communicating internally as well as externally with relevant parties and experts. Although it is difficult to respond to all types of incidents, the company has formulated a rough response policy based on a certain degree of categorization of major AI incidents that can be assumed in light of the content of its AI services. In addition, the company periodically conducts preliminary exercises to confirm the feasibility of the response policy.

Practical Example ii: Establishing a system to prepare for AI incidents through the involvement of external experts.
As a large company that develops and provides AI systems and services, we have established a contact point, assigned a director in charge, and established a system for communication and coordination with the risk management, legal, public relations, and crisis management departments, as well as with external parties and experts, in order to promptly respond to an AI incident. A system is also in place to contact outside parties and experts.

In addition, multiple patterns of possible AI incidents are assumed, and the possible legal liabilities are organized in advance by consulting with experts, and risk assessment is conducted based on these patterns. Since various types of damage may occur, including bodily injury, property damage, invasion of privacy, property damage, etc., it is useful to organize in advance the legal responsibilities of each entity and off-business user for each type of damage. In addition, as a factor to be considered specific to AI systems and services, it is also kept in mind that the causes of outputting abnormal results are diverse (e.g., anomalies in algorithms, authenticity of training data, bias in training data, etc.) and that unexpected effects are likely to occur. Efforts are made to regularly update technical and operational mechanisms to reduce the impact on business even in the event of unexpected situations.

Practical example iii: Establishment of a system to prepare for AI incidents through the inclusion of AI incidents in the BCP

The Company has formulated a company-wide BCP (Business Continuity Plan), but since there is a possibility that business continuity may be disrupted if the AI system operated by the Company is shut down, the Company has decided to include an AI incident as one of the triggers of the BCP, and has formulated a plan for initial response and business continuity in preparation for the case of a total or partial shutdown of the AI system. Therefore, the company has decided to include an AI incident as one of the triggers for triggering the BCP, and has formulated a plan for initial response and business continuity in case all or part of the AI system stops. Recognizing that merely formulating a plan is meaningless and that failure to implement the plan in a contingency would be a major risk, an exercise to put the plan into practice is conducted at least once a year.

4. Operation

Action Objective 4-1 [Ensure accountability of AI management system operation status]: Each entity is expected to fulfill transparency and accountability to relevant stakeholders regarding the operation of the AI management system under the leadership of management, for example, by recording the implementation status of the deviation evaluation process of Action Objective 3-1.

[Points of practice].

Each entity is expected to work under the leadership of management to

- To the extent appropriate and reasonable, make the operation of the AI management system accountable to relevant stakeholders.

In order to increase the accountability of the operational status of the AI management system, the following efforts are useful.

- Documentation of the implementation of the deviation assessment process for Action Objective 3-1
- Maintain records of internal/external meetings regarding the development, provision, and use of AI systems and services (ensuring that they are accessible to parties other than the person in charge)
- Conduct in-house training on AI

As for recording the implementation of the deviation evaluation process, it is also useful to create your own checklist for deviation evaluation, and to check and record the results based on this checklist.

- It is also useful to refer to the checklist in Appendix 7 (attached document) and customize it for your consideration.

In order to be as accurate and understandable as possible for the purpose of explanation to other departments and external parties, the contents of overseas documents, etc., are also helpful.

For example, NIST, "Four Principles of Explainable AI,"[55] , explains the four principles of explainable AI and the five types of explanations.

[Practical example].

[Practice i: Ensure accountability through thorough record keeping and inspection.

Since having data and information in "operations" leads to decision-making for improvement, we believe that the key to improvement through re-analysis and assessment of the environment and risks is in "operations."

The Company places great importance on keeping records not only for AI governance practices but also for further improvement, and it is natural for the Company to keep records on system design. For example, we record deviation assessments in individual AI system development projects, prepare implementation summaries when training on AI is conducted, and keep minutes of internal meetings and meetings with other entities regarding the development and operation of AI systems and services, making them accessible to all parties except those in charge.

Since we are a relatively large company, we have not experienced any difficulties with action goals related to general corporate governance. However, due to the organizational differentiation within the company, we are concerned that AI, a relatively new technology,

[55] NIST, "Four Principles of Explainable Artificial Intelligence (Draft)" (August 2020), https://www.nist.gov/publications /four-principles-explainable-artificial-intelligence
September 2021. The five types are still described only in the Draft.
NIST, "Four Principles of Explainable Artificial Intelligence " (September 2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf

may create gaps in expertise and understanding between departments, which may affect inter-organizational cooperation. For example, with regard to the inquiry desk established in accordance with Action Goal 3-1-2, we are afraid that the person in charge of inquiries may not understand the technical details, which may delay the awareness of serious incidents. In accordance with Action Goal 3-2, efforts are being made to improve employee literacy, and for the time being, inquiries from external parties will be proactively reported to management not only in outline but also in detail.

With regard to the recording of the implementation of the deviation evaluation process in Action Objective 3-1, etc., for the purpose of explanation to other departments and external parties, efforts are made to make the record as accurate and understandable to others as possible, so that they can be aware of the limitations of the explanation.

Practical example ii: Recording using a checklist for small businesses

We are a small company developing AI systems. The technical director is aware of all projects, is very knowledgeable about AI, including programming and reading papers himself, and has a strong interest in AI ethics issues. Therefore, the Company believes that gaps in expertise between departments will not be a problem. On the other hand, due to the high level of expertise of the people involved in the project, they tend to assume that the action goals have been achieved without having to check each and every one of them. For this reason, we have devised a deviation evaluation checklist to be attached to the project progress report reports, so that the technical officers can listen to the reports as necessary.

In addition, we have analyzed that since we are a highly specialized group, there is a tendency for us to be out of step with public perception. Therefore, while checking the status of operations, we are trying to be aware of social acceptance by regularly sharing the status obtained from daily information gathering and exchange of opinions in accordance with Action Objective 3-3-2.

[Practical example iii. Use of checklists throughout the AI lifecycle].

We are a company that develops and provides AI systems and services. We are working to prevent risks before they occur throughout the AI lifecycle by using a checklist.

The checklists are not created from scratch by the companies themselves, but are created by customizing them in their own way, using the "Appendix 7 (Appendix) Checklist" of these Guidelines as a starting point. While some checklist items can be handled by each entity alone, others require collaboration among the entities. The checklist was customized in cooperation with other departments within the company, and was also discussed with the client AI users to create a checklist in the company's own style, taking into account the entire AI lifecycle.

In addition, since a checklist is likely to become a mere skeleton if it is blindly enlarged, we pay attention to the number of items on the checklist, remove items that have become common knowledge within the company, and replace them with the latest items, thereby refining the items to be checked from time to time.

Action Objective 4-2 [Ensure accountable status of individual AI system operations]:
Under the leadership of management, each entity is expected to monitor the status of provisional and full-scale operation of individual AI systems and services, and record the results while implementing PDCA cycle, in order to continuously evaluate deviations in provisional and full-scale operation of AI systems and services. The entity developing the AI system is expected to support such monitoring by the entity providing and using the AI system.

[Points of practice].

Each entity is expected to work under the leadership of management to

- Monitor the status of AI operations of each entity and record the results while implementing the PDCA cycle
- If it is difficult for each entity to respond independently, cooperation among the entities

Specifically, it may be useful for each entity to collaborate in the following cases

- AI developer-driven settings for automatic logging of inputs and outputs that significantly affect performance
- Explanation of specific monitoring methods for AI providers by the AI developer entity
- Discussion of the need for re-training based on output from AI systems and services
- Alignment of expectations for AI systems and services between AI developers and AI providers

[Practical example].
[Practical example i: Coordinated logging between entities].
We are a company that operates AI systems and services and provides said systems and services to AI users. we have commissioned AI developers to develop AI systems and services, from the content of the dataset to checking the behavior of the AI models to ensure that they address not only accuracy but also fairness, We have been briefed by the AI development staff. This developer told us that the AI system/service needs to be maintained to ensure accuracy and fairness in the event of any difference between the image of users assumed at the time of development and the actual image of users.
Since no employee in our company has enough knowledge to interpret the code of the AI system/service, we asked the AI developer to automatically log the inputs and outputs that significantly affect performance, as well as teach us how to monitor the system. Subsequently, as part of Action Objective 3-1, we created a checklist and defined a management method for maintaining performance. Currently, this management method is used to continuously monitor and keep records.

[Practical example ii: Notification of timing of re-study in collaboration with other entities].
The Company is a developer of AI systems/services provided by other companies. Although the Company does not legally own the AI system/service, it has a certain responsibility to operate the AI system/service of others through maintenance contracts, and thus has an aspect of an AI provider. Under such circumstances, the cooperation of the company that routinely operates the AI system/service (AI provider) is indispensable in monitoring to maintain the performance of the AI system/service. In fact, this AI provider is supposed to record the output from the AI system/service, determine from the output any significant deterioration in quality, check the actual situation, and report to us on the need for re-study, and the AI provider is also supposed to participate in subsequent meetings regarding the need for re-study.
The reason why AI providers can decide when to relearn is that they themselves have a good understanding of what they specifically want from AI systems/services and what they specifically can do. It is important for AI developers to understand the expectations of AI system/services from AI providers, and to explain what they can do until AI providers themselves are satisfied. As stated in the AI System Quality Assurance Guidelines, it is important that "quality assurance engineers, teams, and organizations work with development and sales to increase customer understanding of AI products." [56]

Action Goal 4-3 [Consider proactive disclosure of AI governance practices]:
Each entity is expected to consider disclosing information on setting AI governance goals, development and operation of AI management systems, etc., by positioning them as non-financial information in the Corporate Governance Code. Even non-listed companies are expected to consider disclosing information on their AI governance activities. And if, as a result of their consideration, they decide not to disclose such information, they are expected to disclose the fact that they do not disclose such information to their stakeholders, along with the reasons for such decision.

[Points of practice].
Each entity is expected to work under the leadership of management to

[56] AI Product Quality Assurance Consortium "AI Product Quality Assurance Guidelines Version 2023.06" (June 2023)

- Consideration of ensuring transparency of information on AI governance, from the company's basic approach to AI to the development and operation of AI management systems, etc.
- If disclosed, consider placing it in the non-financial information section of the Corporate Governance Code
- If not disclosed, disclose that fact to stakeholders, along with the reasons for not disclosing it.

Specifically, information about AI that is expected to be disclosed may include, for example
Publicizing these information externally is expected to foster a sense of trust, increase recognition, and raise awareness both internally and externally.

- Basic approach/AI policy for your company's AI
- Company's Efforts on AI Ethics
- In-house AI Governance

[Practical example].
Practical example i: Disclosure of AI governance goals via website, etc.

As a small company developing AI systems, we believe that the development of AI systems must be supported by a deep understanding of society, not just a technical endeavor, and we prioritize internal instillation of this mindset rather than explicitly setting AI governance goals. Customers and shareholders support this stance. Of course, we believe it is important to respect the "common guiding principles" of these Guidelines, but it is the understanding of the philosophy and other principles behind them that is important.

As a privately held company, we are not subject to the Corporate Governance Code, but we are actively communicating the aforementioned approach to AI through our website and other means. Our potential customers and non-business users of our AI systems and services perceive our AI systems and services as socio-technical tools rather than technical tools, which differentiates us from our competitors.

Example of Practice ii: Consideration for inclusion in non-financial information

As a publicly listed company that develops AI systems, the appropriate development of AI is an important theme for us, and we have already set our own AI policy and completed the development of a system to achieve it. The company has also announced these activities on its website and made press releases. On the other hand, we have considered issuing a strong message from management regarding these activities, but have not yet reached the point of issuing such a message because our AI-related business does not directly affect our medium- to long-term earnings at this time.

In this context, we recently received a corporate governance questionnaire from an institutional investor, which included a question on how the company is dealing with AI ethics. If such a questionnaire reflects investors' intentions toward the medium- to long-term development of a company, it suggests that AI ethics is also necessary information for judging the sound development of a company. In the future, we plan to once again consider proactive information dissemination from management, including the inclusion of AI ethics initiatives in the integrated report.

5. Evaluation

Action Goal 5-1 [Verify the functioning of the AI management system]: Each entity, under the leadership of management, is expected to have a person with relevant expertise independent from the design and operation of the AI management system to determine whether the AI management system, including the deviation evaluation process, is appropriately designed and properly operated in light of the AI Governance Goals, i.e., Action Goals 3 and 4. Through practice, it is expected that the AI Management System will be required to evaluate and continuously improve whether it is functioning appropriately to achieve the AI Governance Goals.

[Points of practice].

Each entity is expected to work under the leadership of management to

- Clearly articulate in management's own words the key points of the evaluation for continuous improvement.
- Assign a person with relevant expertise independent of the design and operation of the AI management system
- Monitoring of the proper functioning of the AI management system by the above mentioned persons
- Continuous improvement based on monitoring results

Persons with relevant expertise independent of the design and operation of the AI management system may specifically include the following

- In case of in-house implementation
  - Internal Audit Department
  - Self-audit of AI management system with AI developers who are not involved in the work to be audited, etc.
- When to use outside resources
  - External audit entities, international organizations, etc.[57]
    - Those with a high level of expertise and the ability to utilize and apply audit experience from other companies.

In each case, it is also important to keep in mind the following points

- In case of in-house implementation
  - Taking measures to increase effectiveness, such as requiring reporting to the department in charge of risk management and the officer in charge of AI governance (the person in charge of audit reporting directly to that officer).
  - Consideration should be given to ensure that the evaluation is not superficial because the internal audit department is not familiar with AI, for example, by assigning a person in the internal audit department who can understand the technical aspects of AI, or by having each department cooperate in audits conducted by the internal audit department.
    - For example, the audit findings are biased toward the operational processes that are easy to see, and there are few findings on the design and development processes, etc.
- When to use outside resources
  - External audit entities are not necessarily knowledgeable about each entity's unique issues and specific circumstances. Therefore, it is important for each entity to voluntarily collect information on social acceptance and engage in dialogue with stakeholders, rather than leaving it up to the external audit entity, etc.
  - A possible case where there is a high need to use external resources could be a situation where there is a need to explain to relevant stakeholders whether the AI management system is functioning properly. In such a case, it is important to clarify in which country, according to which management and evaluation standards, and in

[57] World Economic Forum, "The Presidio Recommendations on Responsible Generative AI" (June 2023).

which scope evaluation and reporting are required, and then to select external resources with the expertise to conduct such evaluation.

The management and audit organization standards[58] are currently under discussion internationally and it is expected to check the trend.

[Practical example].
Example of practice i: Monitoring through the internal audit department

Minebea has an independent internal audit department that has been auditing the operation of internal rules and regulations since before the introduction of the AI management system. If any inappropriate operation or malfunction is found, the internal auditor asks the department to make improvements and shares best practices of other departments, if any.

Social acceptance of AI systems and services is changing. MSI believes that improvements that keep pace with social acceptance are important, and internal audits are conducted with reference to environmental and risk analysis, focusing on areas where social expectations are high and where the number of incidents is high. In order to obtain cooperation from each department for improvement, the company does not uniformly conduct strict conformity assessment to internal rules, etc. for all areas, but selects areas with high risk. Communicating the reasons for selection makes it easier to obtain cooperation from each department.

Example of practice ii: Monitoring using self-audits

As a small company developing AI systems, we do not have an internal audit department to evaluate our AI management system, but conduct self-audit with the addition of people in the development department who are not directly involved in the AI management system. Since the first line of auditing, self-audit, tends to be too lenient on the company's own staff, the results of self-audit are reported to the audit staff directly under the director in charge of AI governance, and the reports are organized and reported to the director in charge of AI governance. We believe that the system is functioning adequately despite its focus on self-auditing. Currently, we are considering holding cross-departmental feedback meetings to share audit results and exchange opinions in order to strengthen the third-party perspective and to communicate that internal audits are for the improvement of AI systems.

Example iii: Combined internal and external monitoring

Although we have an internal audit department, we have decided to utilize an external audit for our AI management system. We expect external audits to have a high level of expertise and to horizontally expand the auditing experience of other companies. Even if the company is proud that it is fully capable of handling AI systems in its own way, there may be blind spots.

External audit services are provided mainly by consulting firms. By having an audit conducted by an external expert, the client can receive advice that utilizes specialized information from both inside and outside the company. The third-party nature and objectivity of the advice provided by external experts is also expected to have the effect of providing smoother feedback to the company.

While this is a benefit, we are concerned about the potential for passivity. External experts are not necessarily knowledgeable about issues specific to each company. In order to make the best use of advice from external experts, it is important to be willing to actively understand the social acceptance of AI, even when relying on external audits.

Action Objective 5-2 [Consideration of stakeholder input]:
Each entity is expected to consider seeking opinions from stakeholders on the AI management system and its operation under the leadership of management. If, as a result of the

[58]Information technology-Artificial intelligence-Management system as a management standard and ISO/IEC42006 as a standard for auditing bodies. Information technology-Artificial intelligence-Requirements for bodies providing audit and certification of The standard for auditing organizations is ISO/IEC 42006 Information technology-Artificial intelligence-Requirements for bodies providing audit and certification of artificial intelligence management systems.

consideration, it is decided not to implement the content of such opinions, it is expected to explain the reasons to stakeholders.

[Points of practice].

Each entity is expected to work under the leadership of management to

- Consider seeking input from stakeholders on the AI management system and its operation.
- If the opinion is not implemented, explain the reasons to stakeholders.

In addition, it is important to build a network through the following efforts so that you can collaborate with stakeholders and be in a position to obtain advice that is relevant to your company's situation on a daily basis.

- In-house training with outside instructors
- Forming a loose network and exchanging information outside of work with people who are interested in AI ethics and quality
- Proactively utilize opportunities for exchanging opinions on AI ethics and quality, conferences, etc.
- Establish an organization including external experts on AI governance and other experts in AI and other fields.

[Practical example].

Practical example i: Study on AI governance through an organization including external experts, etc.] [Practical example ii: Study on AI governance through an organization including external experts, etc.

The Corporate Governance Code's chapter on "Appropriate Collaboration with Stakeholders Other Than Shareholders" summarizes that it is important to strive for appropriate collaboration with stakeholders, including employees, customers, business partners, creditors, and local communities, and that the board of directors and management in particular should exercise leadership in fostering a corporate culture and climate that respects the rights and positions of these stakeholders and sound ethics in business activities. In particular, the report summarizes that it is important for the board of directors and management to exercise leadership in fostering a corporate culture and climate that respects the rights and positions of these stakeholders and sound business ethics. In addition, given the growing interest in the appropriate development, provision, and use of AI systems and services, listed as well as non-listed companies may be required to collaborate with their stakeholders in evaluating and reviewing their AI governance and AI management systems.

While the Company believes that the initial setting of the AI policy and the establishment of a system to achieve the AI policy should be done by the companies themselves, and that subsequent improvements should also be made proactively by the companies themselves, the Company also emphasizes collaboration with stakeholders to understand "how society sees the Company". The Company has already established an AI policy and publicly announced the meaning of the AI policy and activities to achieve the AI policy. However, we believe it is necessary to know "how we are seen by society" and to ensure objective ethics. For the purpose of engaging in dialogue with stakeholders, we have decided to establish an organization including external experts on AI governance, which is composed of specialists in AI and other fields, not only experts in AI technology but also Experts in legal, environmental, and consumer issues have also been invited. Since it is not sufficient just to receive general remarks, the committee is devised to present the Company's specific issues to gain deep insights.

Practical example ii: Consideration using a forum for the exchange of opinions

While we tend to focus on "visible measures" such as the establishment of a committee of outside experts, as in Practice i, we believe that such a forum is not all there is. What is important is to be loosely connected to a network of people interested in AI ethics and quality and to be part of this information exchange network. We encourage our management to be active in exchanging ideas on AI ethics and quality, and to actively take on the role of speaker at conferences and other events. Of course, we include such activities in our performance evaluations.

We hear concerns that this approach does not gather opinions. One possible reason for this concern is that Japanese people do not speak frankly at opinion exchanges and conferences. However, people of the so-called "active sonar type," who elicit others' opinions by expressing their own, know that there are people who will personally give their opinions after an opinion exchange or conference. They believe that obtaining opinions from such people is what is important.

Following the connections of this management team, we once held an in-house training session with an external lecturer. In this training, in addition to explaining our AI governance initiatives to employees engaged in AI-related work, we asked the external lecturer to evaluate our company's initiatives. The external lecturer, who exchanges opinions with our management on a daily basis, was able to provide advice that was relevant to our company's situation, and the training was well received by the participants.

Because of this situation, we are considering the establishment of an outside expert committee, but do not see the need for one at this time.

6. Re-analysis of environment and risk

Action Objective 6-1 [Timely Re-implementation of Action Objectives 1-1 to 1-3]:
Under the leadership of management, each entity is expected to promptly identify changes in the external environment, such as the emergence of new technologies and changes in regulations and other social systems, with regard to Action Objectives 1-1 through 1-3, and to reevaluate, update its understanding, and acquire new perspectives in a timely manner, in order to improve or reconstruct AI systems and improve operations based on such changes. It is expected that the AI system will be improved or reconstructed and its operation will be improved based on these changes. In implementing Action Goal 5-2, it is expected to consider not only the existing AI management system and its operation, but also to obtain external opinions for reviewing the entire AI governance, including environmental and risk analysis, in line with agile governance, which is also emphasized in this guideline.

[Points of practice].

Each entity is expected to work under the leadership of management to

- To understand changes in the external environment, such as the emergence of new technologies, technological innovations related to AI, and changes in regulations and other social institutions
- Timely re-evaluation, updating of understanding, acquisition of new perspectives, etc., and improvement, restructuring, operational changes, etc., of the AI system in line with such re-evaluation, updating of understanding, acquisition of new perspectives, etc.
- To embed the concept of AI governance into the culture of the organization.

It is also important to obtain external information on social trends through such means as holding periodic meetings with outside experts.

Timely reanalysis varies from entity to entity, but apart from periodic (quarterly/semiannual/annual, etc.) implementation, the following timings are also candidates

- In the event of a serious "near-miss
- When a serious AI incident occurs at another company
- When society's attention is focused on a particular AI technology or AI incident
- Socially, when the regulatory environment changes

For example, the following innovations are useful in establishing a system to recognize the occurrence of serious "hiyari-hatto" incidents.

- Establishment of a system to facilitate employee reporting of "Hiyari-Hatto" (near-misses)
  - Introduction of an anonymous reporting system, introduction of a reward system for those who report near-misses, educational activities, etc.
- Establish a regular risk assessment and monitoring system

In order to make the system and operations related to AI governance function, the concept of AI governance should permeate the entire organization and take root as a culture. To this end, it is important for each entity's members to be aware of their own roles in AI governance and to have a sense of party to the overall optimization so as not to fall into partial optimization. Examples of efforts to foster a culture within each entity include the following

- Introduction of a personnel evaluation system that recognizes steady day-to-day AI governance transfer activities such as cross-organizational consortiums and community activities.
- Training at the time of new employee assignment and at the time of transfer
- References to stance on AI governance in employee-mandated standards of conduct, brochures, etc.
- Regular e-learning, training and education, etc.

[Practical example].

[Practice i: Re-analysis in accordance with reporting opportunities to management].

We regularly analyze the environment and risks and report to management, except in the event of a major "near-miss," a significant increase in public attention to a particular AI incident, or a change in the regulatory environment, etc. The appropriate development, provision, and use of AI systems and services While the discussion around the appropriate development, provision, and use of AI systems and services is very active, the emphasis is on preventing AI governance fatigue through agile reanalysis and on identifying major trends in an agile manner. Reporting opportunities to management are good opportunities to look at major trends.

[Practical example ii: Re-analysis in accordance with the implementation of the conference body including external experts, etc. on AI governance].

Although the Company regularly analyzes the environment and risks as described in Practice i, because there are some overlapping elements in the verification of AI governance and AI management systems, the Company has included in the agenda of a regularly convened organization that includes invited external experts on AI governance, the benefits/risks that AI systems/services may bring and the social acceptance of the development and provision of AI systems/services, to obtain major trends on these issues from external experts. The agenda of the regularly invited external experts on AI governance includes the possible benefits/risks and social acceptance of the development and provision of AI systems and services, in order to obtain major trends on these issues from the external experts.

B. Examples of Actual Efforts to Establish AI Governance

Examples of businesses promoting AI governance will be discussed.

Column 4 NEC Group's AI Governance Initiatives

In 2018, NEC established the Digital Trust Promotion Management Department as an organization responsible for formulating and promoting company-wide strategies to ensure that business activities related to the use of AI respect human rights, and in 2019, formulated the NEC Group Policy on AI and Human Rights (hereinafter "Company-wide Policy"). As a governance structure, a Chief Digital Officer (CDO: Chief Digital Officer) has been appointed to clarify the relationship with the Risk Compliance Committee and the Board of Directors, and the "Digital Trust Advisory Council" of external experts has been established to proactively collaborate with external parties, etc. The company has also established the Digital Trust Advisory Council, an external expert panel, and is actively collaborating with external parties to address AI governance as part of its management agenda (see Figure 8: Promotion Structure of AI Governance).

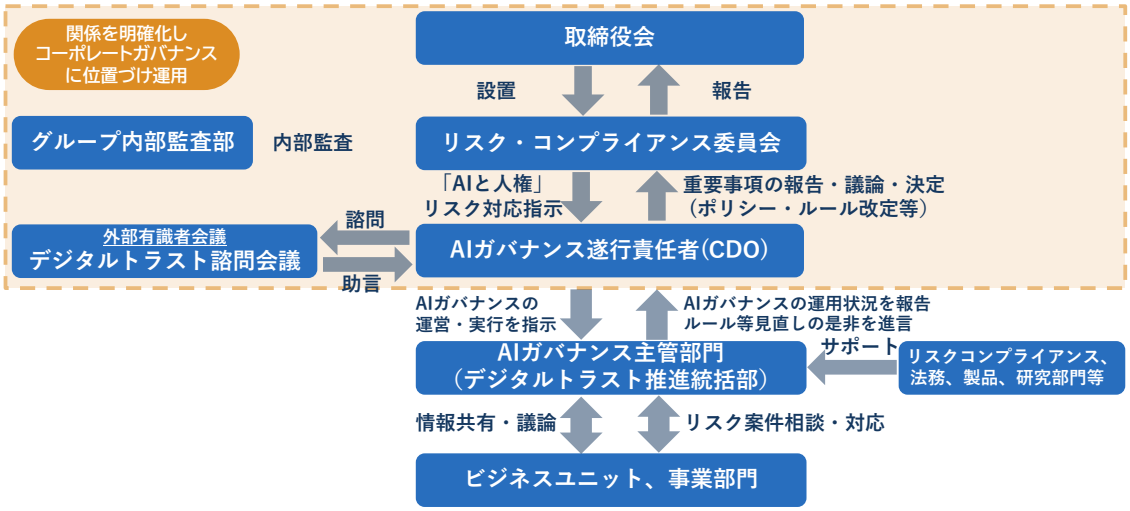

Figure 5 . **AI Governance Promotion Structure (To be translated later)**

The company-wide policy was reviewed by the Digital Trust Promotion and Management Department based on domestic and international principles and the company's vision, values, and business activities, and was discussed with various internal and external stakeholders, including relevant departments within the company such as R&D, sustainability, risk management, marketing, and business divisions, as well as external experts, NPOs, and consumers The policy was formulated in April 2019 through a process of This policy was formulated to give top priority to privacy considerations and respect for human rights in the use of AI, and consists of seven items: fairness, privacy, transparency, accountability, proper use, AI development and human resource development, and multi-stakeholder dialogue. It consists of.

In order to put the company-wide policy into practice, the Digital Trust Promotion & Management Department is taking the lead in developing internal systems and training employees. Specifically, the company has established company-wide regulations that stipulate governance systems and basic matters to be observed, guidelines and manuals that stipulate matters to be addressed and operational flow, and risk check sheets, and has put in place a framework for checking risks and implementing countermeasures for AI utilization in each phase, starting from the planning and proposal phase.

Web-based training for all employees and training for AI business personnel and management are also conducted, with outside experts as lecturers, to promote understanding of the latest market trends and case studies (see Figure 9. Overall picture of AI governance).

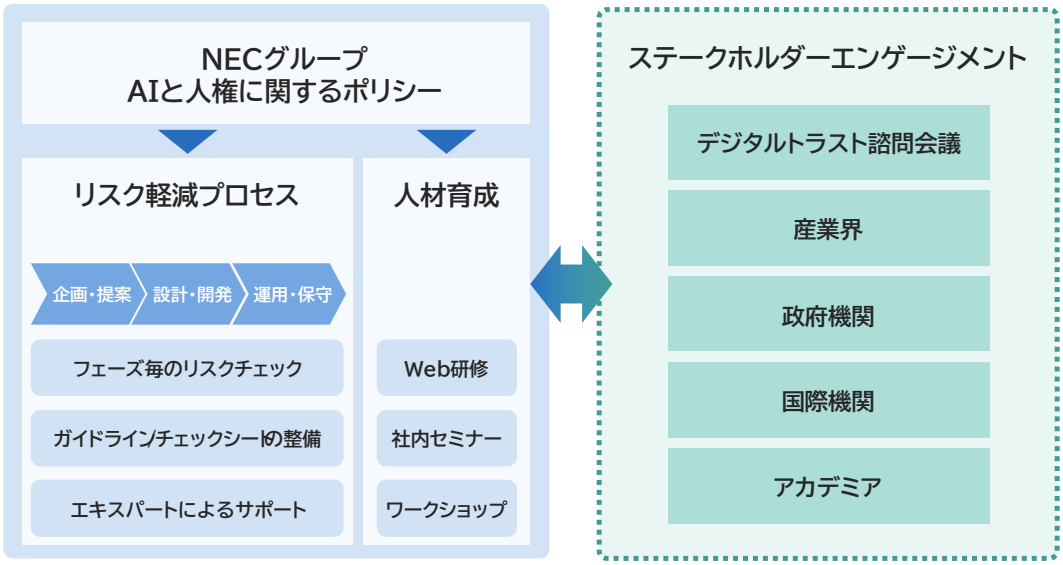

Figure 6．Overall picture of AI governance **(To be translated later)**

In these efforts, five levels of "maturity" are defined for each of the 21 action goals listed in the Ministry of Economy, Trade and Industry's "Governance Guidelines for AI Principle Implementation" (hereinafter referred to as the former "AI Governance Guidelines") to visualize the current status of AI governance, which is used to set action items to achieve the goals and to manage progress. This information is used to set up and manage progress (See Figure 10. How to use the former "AI Governance Guidelines"). In addition, based on the concept of agile governance in the former AI Governance Guidelines, the company is flexibly responding to changes in the social environment and reviewing internal rules and operations. In 2023, the company will establish rules for internal use of generative AI (large-scale language models).

成熟度 行動目標	Lv.1 Performed 個々人による 単発的な実施	Lv.2 Managed ポリシーに従った 反復的実施	Lv.3 Defined 統一された標準的 プロセスの確立	Lv.4 Measured 定量的な 評価の実施	Lv.5 Optimized フィードバックに基 づく継続的最適化
1-1. AIシステムから得られる正のインパクトだけではなく…	国内及び同業他社に関するAI利用のガイドライン…				
1-2. 本格的なAIの提供に先立ち、直接的なステークホルダー…	政府、市民団体等が公表している消費者アンケートや、AI利用に…				
…	…				
6-1. 行動目標1－1から1－3について、適時に再評価…	重大な「ヒヤリ・ハット」が生じた場合、特定のインシデントへ…				

Figure 10．Use of the former AI Governance Guidelines **(To be translated later)**

Column 5 Toshiba Group's AI Governance Initiatives

The company has established an "AI-CoE Project Team" to lead group-wide AI measures in 2022, and formulated an "AI Governance Statement" that concretizes the group's management philosophy system in terms of AI utilization. While referring to the former "AI Governance Guideline," the AI-CoE project team is building "AI Governance" based on this statement. Within this framework, the "AI-CoE Project Team" is leading the activities to promote "AI Governance" by forming a working group with experts in various fields such as privacy, security, and legal affairs, as well as representatives from the business side.

Specifically, in addition to visualization and promotion of utilization of the Group's AI technology assets by establishing an AI technology catalog and AI human resource development through its own training program, the Group is working on a system to maintain the quality of its AI systems through the development of MLOps (a system to manage the life cycle of machine learning models) and an AI quality assurance system. (See Figure 11. Outline of the Group's AI Governance).

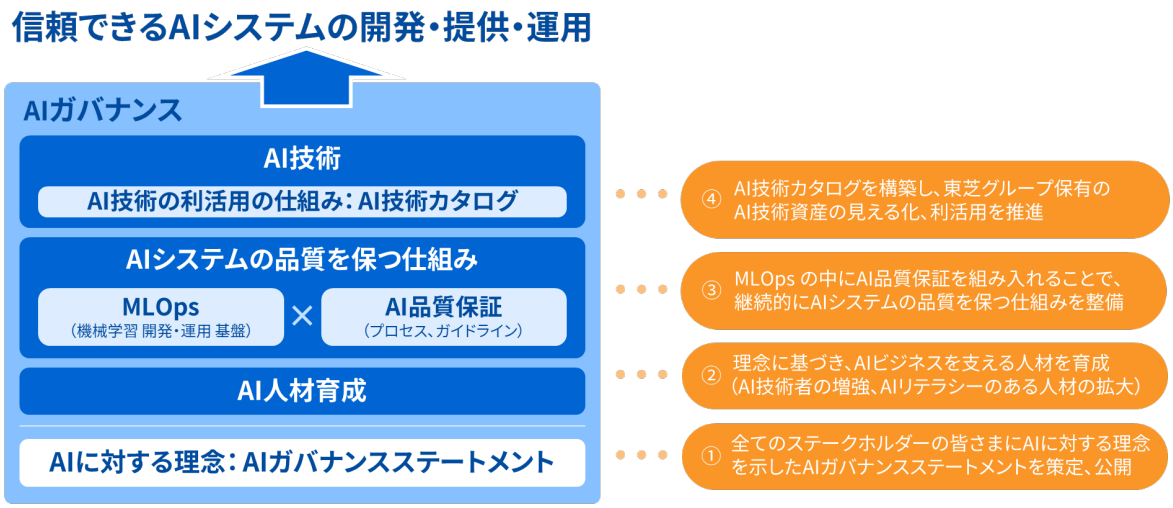

Figure 7 . Overview of the Group's AI Governance **(To be translated later)**

This "AI Governance Statement" is intended to reflect Toshiba Group's management philosophy system and to clearly state its philosophy on AI. It consists of seven elements: "Respect for People," "Ensuring Safety and Security," "Thorough Compliance," "Development of AI and Human Resources," "Realization of a Sustainable Society," "Emphasis on Fairness," and " 7 elements of "Emphasis on transparency and accountability".

Based on this statement, a system to maintain the quality of AI systems has been established based on the two axes of "AI Quality Assurance" and "MLOps. In "AI Quality Assurance," we have established "AI Quality Assurance Guideline," which outlines the concept and issues to be addressed in the development of AI systems, and the "AI Quality Assurance Process," which identifies necessary tasks and deliverables to be created based on this guideline, and organizes the process without any omissions. In addition, through the "Quality Card," AI quality assurance, which tends to be from the developer's perspective, is evaluated from the user's perspective, and efforts are made to visualize AI quality.

In MLOps, business and machine learning experts, system developers, and system operators work as a unified team to continuously improve AI systems to prevent performance degradation due to environmental changes after operation begins. By coordinating these efforts, the company is practicing the development, provision, and operation of reliable AI systems.

By initiating these AI governance initiatives, not only AI experts (engineers) but also the entire Toshiba Group has improved the literacy required for the development, provision, and operation of AI systems (i.e., increased awareness of risks as well as opportunities for AI use).

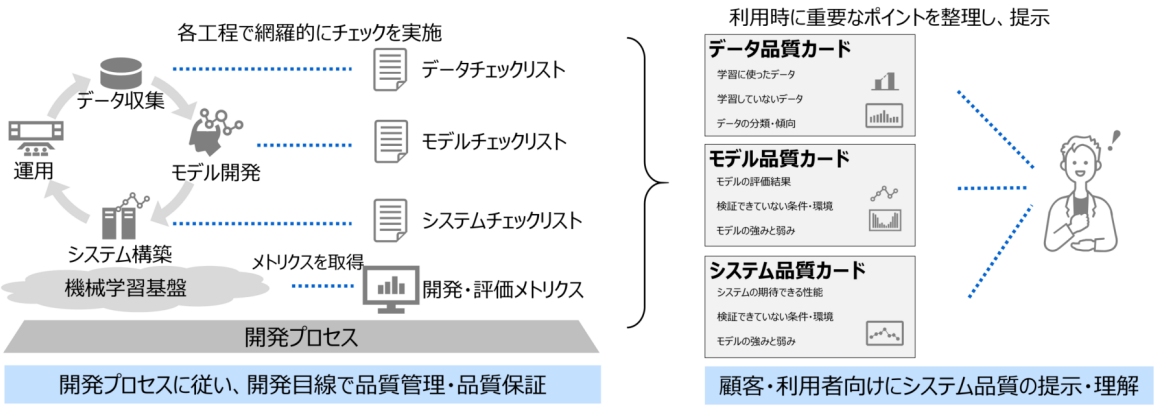

Figure 8 . Flow of AI Quality Assurance Guidelines and Quality Card Utilization **(To be translated later)**

column (e.g. in a magazine) 6 Panasonic Group's AI Governance Initiatives

In 2019, the company established an AI Ethics Committee within the former Panasonic Corporation to formulate AI Ethical Principles to be observed within the company; in 2022, the committee was reorganized into the Panasonic Group AI Ethics Committee as an organization to operate the AI Ethical Principles across the Group, and in the same year, the Panasonic The AI Ethical Principles of the Panasonic Group" was published in the same year. Currently, the AI Ethics Committee is playing a central role in developing and utilizing the AI Ethics Check System, which will be operational from 2022, and in providing AI ethics education for all employees (see Figure 13. AI Governance Structure Overview).

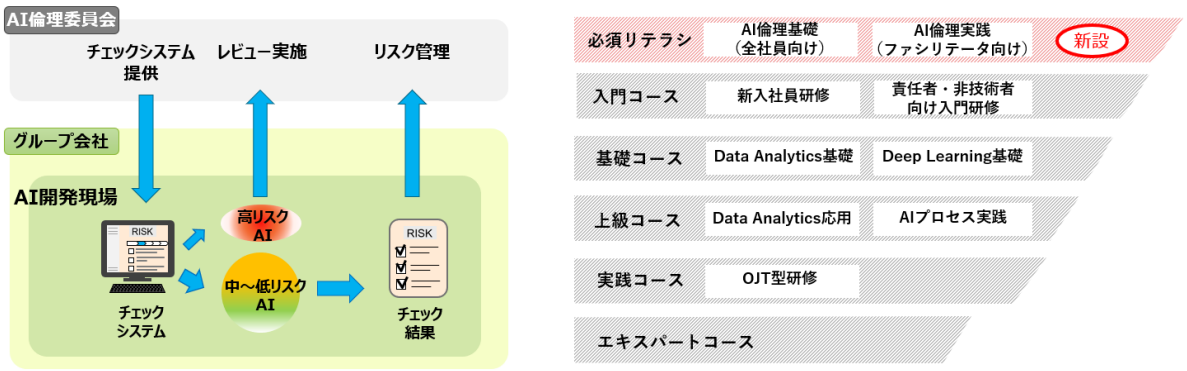

Figure 9 . AI Governance Structure Overview **(To be translated later)**

The AI Ethics Committee was established within Panasonic Holdings, Inc. to publicize the AI Ethical Principles and to implement activities to earn the trust of users and society in a wide range of business areas. Specifically, one or more AI ethics officers have been selected from each of the Group's operating companies to form a group-wide AI ethics promotion system together with the legal, intellectual property, information system/security, and quality divisions (see Figure 14. AI Ethics Committee Structure). In order to respond to the Panasonic Group's diverse business fields, each AI Ethics Officer promotes AI ethics activities within the group of operating companies, and the AI Ethics Committee provides support for these activities.

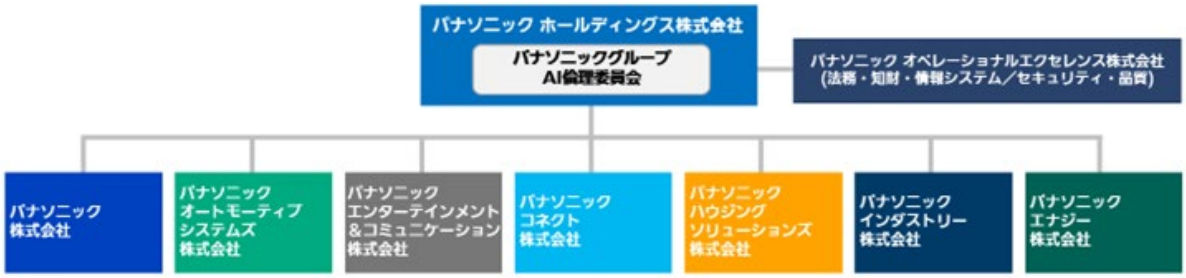

Figure. 10 . AI Ethics Committee Structure **(To be translated later)**

As one of the initiatives of the AI Ethics Committee, an "AI Ethics Check System" has been developed. The purpose of this system is to efficiently and effectively check AI ethical risk while preventing an increase in on-site workload and hindering innovation in the Group's diverse and wide-ranging use of AI. The system is designed to generate a checklist sufficient for the characteristics of products and services, and to confirm that AI under development does not deviate from ethical AI principles. The system also provides enhanced explanations for each check item, as well as information, technologies, and tools for countermeasures, to enable the field to proactively check AI ethics and promote improvements. The results of the

self-checks are collected, analyzed by the AI Ethics Committee, and reflected in the activities. The check items are based on the former "AI Governance Guidelines" of the Ministry of Economy, Trade and Industry (METI), and the first edition was prepared based on domestic and foreign guidelines.

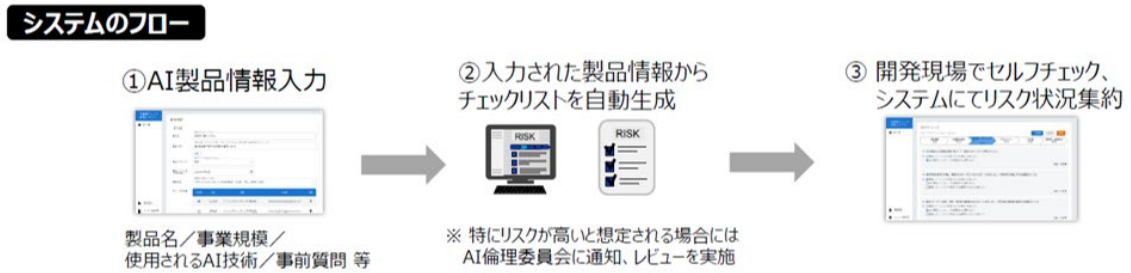

Figure 11 . AI Ethics Checking System **(To be translated later)**

column (e.g. in a magazine) 7 Fujitsu Group's AI Governance Initiatives

As an AI developer and provider, the company is committed to eliminating concerns and unforeseen inconveniences related to AI and creating a sustainable society through appropriate use of AI technology. In addition to active participation in international AI ethics discussions, we will promote advanced internal governance initiatives such as the "AI Ethics Check" and the establishment of AI ethics officers in overseas regions, while also focusing on efforts to promote AI ethics outside the company by introducing AI governance initiatives and publishing guidelines for the use of AI generated by the company.

With reference to the five principles proposed by AI4People, a European consortium that joined in 2018, the Fujitsu Group AI Commitment was formulated in 2019, and to further put it into practice, specific criteria and procedures were developed according to the way AI is used (Figure 16. Fujitsu Group Commitment Fujitsu Group Commitment). In addition, in order to obtain an objective evaluation of AI governance efforts, the Fujitsu Group AI Ethics External Committee was established, and external experts with an emphasis on diversity, including biomedical science, ecology, law, SDGs, and consumer issues, as well as AI technology, have been invited as its members. The committee, in which the president and other members of the management participate as observers, compiles active discussions into recommendations that are shared with the board of directors, thereby incorporating AI ethics into corporate governance as a "key issue in corporate management.

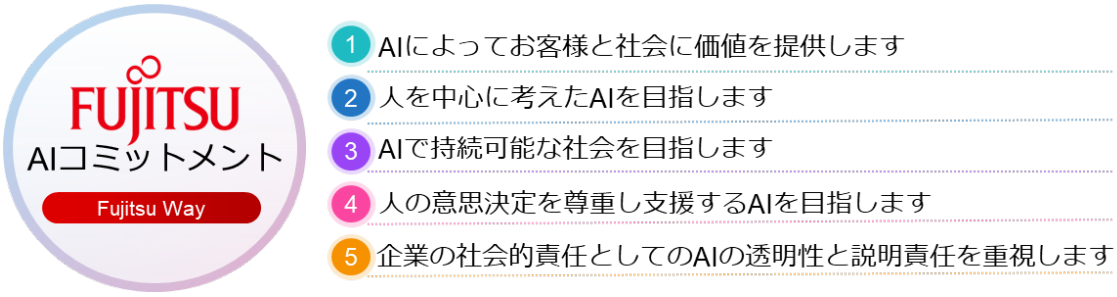

Figure 12 . Fujitsu Group Commitment **(To be translated later)**

As a result of the institutionalization of education on AI ethics starting in 2020, the level of employee awareness has increased dramatically, and as a consulting service we are able to advise AI user companies on ethical perspectives.

In 2022, recognizing that AI ethics is a group-wide management issue, the AI Ethics Governance Office[59] was established directly under the company (Corporate Division) as an organization to lead the AI ethics strategy. In doing so, the office has recruited people from various parts of the group with a wide range of experience, including those with development and sales backgrounds, and has also created an open environment where individual opinions are respected so that the generation of digital natives can play an active role. In this office, frank exchanges of opinions and suggestions are routinely held, and the various AI ethics penetration measures generated from these discussions are promoted throughout the group (see Figure 17. AI Ethics Governance Structure).

[59] For more information, please refer to the white paper "Recommendations by the Fujitsu Group AI Ethics External Committee and Examples of Fujitsu's Practices" posted on Fujitsu's dedicated website on AI Ethical Governance. https://www.fujitsu.com/jp/about/research/technology/ai/aiethics/

Appendix 2. "Part 2 E. Building AI Governance " Related
column (e.g. in a magazine) 7 Fujitsu Group's AI Governance Initiatives

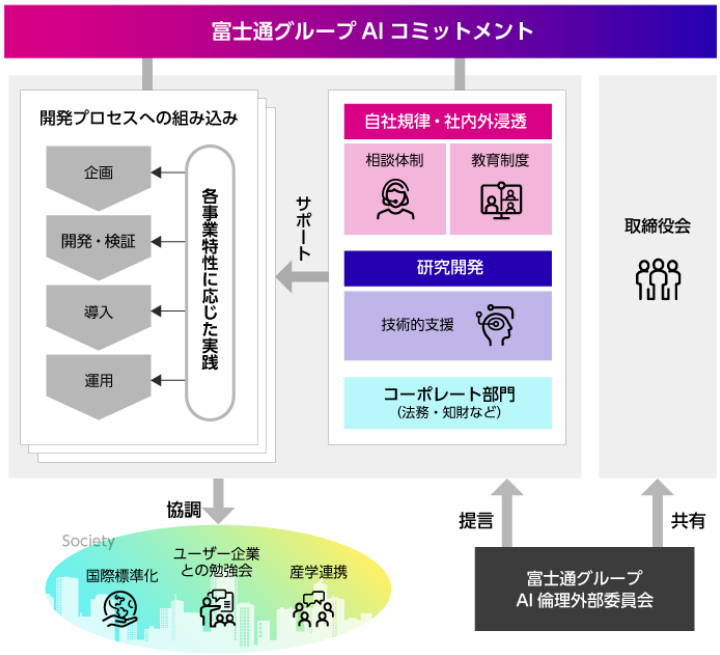

Figure 13 . AI Ethical Governance Structure **(To be translated later)**

Furthermore, in 2023, the "AI Ethics Check" will be made mandatory for all business transactions in Japan, and any case involving ethical issues will be subject to a decision on promotion and improvement after a meeting of legal, R&D, DE&I, business divisions, etc., to ensure thorough control that goes beyond AI-related quality assurance and security perspectives. Overseas Overseas, in addition to the ethical review at the head office, a person in charge of AI ethics is assigned to each region to conduct ethical checks at the time of AI implementation in the region.

In addition, through the free release of the "AI Ethics Impact Assessment," we are working to promote the development and provision of safe, secure, and reliable AI both internally and externally. The "AI Ethical Impact Assessment" was developed to assess the ethical impact of AI on people and society by extracting items relevant to developers and operators of AI systems in compliance with various domestic and international AI ethical guidelines, including guidelines published by the Cabinet Office, Ministry of Internal Affairs and Communications, and Ministry of Economy, Trade and Industry, as well as OECD, EU and US guidelines. The guidelines were formulated in order to evaluate the ethical impact of AI on people and society. In addition to this public release, study sessions with user companies, industry-academia collaboration, and standardization activities are being conducted to promote the spread of AI ethical initiatives throughout society (see Figure 18. Outline of AI Ethical Impact Assessment).

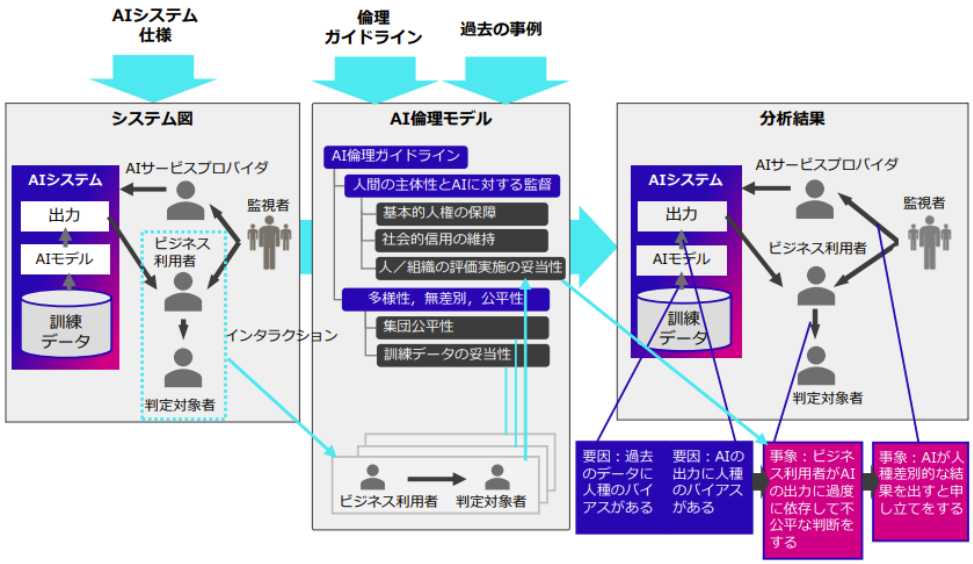

Figure 14 . Overview of AI Ethical Impact Assessment **(To be translated later)**

Appendix 3. for AI Developers

In this chapter, first, "Points" and "Specific methods" of the contents described in "Part 3: Matters Concerning AI Developers" of this volume will be explained. Then, among the "C. Common Guidelines" in "Part 2: Society to be Aimed by AI and Matters to be Tackled by Each Entity" of this volume, specific methods that AI developers should be particularly aware of are explained.

Note that the "specific methods" described here are only examples. Some of them apply to both conventional AI and generative AI, while others apply only to one or the other. In considering specifics, it is important to take into account the degree and probability of risk posed by the AI to be developed, the technological characteristics, and the resource constraints of each actor.

Also, actors developing the most advanced AI systems should should take into consideration the "Hiroshima Process International Guidelines for Organizations Developing Advanced AI Systems" established in the Hiroshima AI Process[60] (described in "D. Guidelines Common to Entities Involved in Advanced AI Systems" in "Part 2.) and "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems"[61] (see C. Matters to be observed in developing advanced AI systems below).

[60] Ministry of Foreign Affairs of Japan, "International Guidelines for the Hiroshima Process for Organizations Developing Advanced AI Systems" (October 2023), https://www.mofa.go.jp/mofaj/files/100573469.pdf
[61] Ministry of Foreign Affairs of Japan, "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems" (October 2023), https://www.mofa.go.jp/mofaj/files/100573472.pdf

A. Explanation of "Part 3: Matters related to AI Developers"

[Description in this volume (reprinted)]
During data preprocessing and learning

D-2) i. Study of appropriate data

- ✧ Through Privacy by Design and other means, ensure that data at the time of learning is collected appropriately and, if it contains confidential information of third parties, personal information, or intellectual property rights that require attention, that it is handled appropriately in accordance with laws and regulations throughout the entire life cycle of AI ("2) Security", " (4) Privacy Protection" and "5) Security Assurance")
- ✧ Implement appropriate safeguards, such as considering the introduction of data management and restriction functions to control access to data before and throughout the study ("2) Safety" and "5) Ensuring Security")

[Point]

Based on the importance of improving the quality of data and models, it is important for AI developers to pay attention to the quality of data used for training AI systems and other purposes.

- Pay attention to the quality (accuracy, completeness, etc.) of data used for AI training, etc., based on the characteristics of the AI to be used and its application.
- In addition, it is expected that standards for accuracy should be established in advance based on the scale of expected infringement, frequency of infringement, technological level, cost to maintain accuracy, etc., since it is assumed that the accuracy of judgments made by AI may be impaired or degraded subsequently. If the accuracy falls below the standard, the data is to be re-trained, paying attention to the quality of the data.
- "Accuracy" here includes whether the AI makes ethically correct judgments (e.g., whether the AI uses violent expressions, does not engage in hate speech, etc.).

[Specific Methods]

- Verify that data does not contain personal data, confidential information, or copyrights or anything related to legally protected rights or interests
  - ➢ Named entity recognition
    - ✧ Name of person, credit card number, etc.
- Implement appropriate handling of personal information, confidential information, copyrights, etc., when it contains items related to rights and legally protected interests
  - ➢ Differential privacy
    - ✧ Adding noise to the data so that the AI developer never knows the actual data
  - ➢ Data Management Console
    - ✧ Provide tools and consoles that allow the person who has provided personal information to determine whether or not to provide personal information, the extent to which he or she may provide personal information, withdraw consent, etc., and to easily track the current status of this information.
  - ➢ data encryption
    - ✧ Use strong encryption algorithms to protect information when transferring or storing data
- Implement measures to ensure that data is appropriate (i.e., accurate, complete, etc.) and secure
  - ➢ Verifying time stamps, etc.
- Implementation of measures to ascertain the source of data to the extent technically feasible and reasonable
  - ➢ Data lineage (construction of a comingled mechanism)
    - ✧ Knowing where the data originally came from, how it was collected, managed, and moved within each actor over time

- ✧ Such data includes the identifier of the service or model that created the content, but need not include user information

[References]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- NIST, "AI Risk Management Framework Playbook" (January 2023).

[Description in this volume (reprinted)]
During data preprocessing and learning

D-3) i. Consideration for bias in data

- ✧ Take reasonable steps to control data quality, noting that training data and AI models can contain biases (including potential biases that do not appear in the training data) due to the training process ("3) Impartiality").
- ✧ Given that bias cannot be completely eliminated from the learning process of training data and AI models, development based on various methods, rather than a single method, will be conducted in parallel as needed ("3) Fairness").

[Point]

AI developers are mindful that the judgments of AI systems may contain biases. AI Developers are also mindful that individuals and groups are not unfairly discriminated against based on the judgements of AI systems.

- Data is only a slice of an event or phenomenon and does not fully reflect the real world. Therefore, note that there is a risk that the dataset may be biased or that a particular community may be under- or over-represented on the dataset. Also, ensure that there is no bias or under- or over-representation in the underlying data set.
- Be mindful in relation to "fairness" as real-world biases and prejudices may be latent in the data set, resulting in the inheritance and reproduction of existing discrimination.

[Specific Methods]

- Before the training
  - ➢ Determination of features not to be used [62]
    - ✧ Not training models on attributes such as race, ethnicity, or gender that could result in bias or discrimination, except in limited cases, such as checking to see if unfair bias is occurring in the AI system
    - ✧ In determining which attributes not to train in the AI model, consider the reasons listed in Article 14.1 of the Constitution (all citizens are equal before the law and shall not be discriminated in political, economic, or social relations because of race, creed, sex, social status, or family origin), as well as attributes mentioned in international human rights rules. take into account
    - ✧ Taking into account the approximate level of data volume required for the intended behavior in order to avoid bias due to too little data for the AI to be trained.
  - ➢ Management and improvement of data quality
    - ✧ Reconstructing the dataset
      - For example, remove some data and adjust annotation content so that the sex ratio of the data is appropriate for the purposes of AI development.
    - ✧ Reviewing of labels
      - Note that labels for training data are often created and assigned by humans during data preprocessing, which introduces (intentional or unintentional) the bias of the person assigning the labels.
    - ✧ Attention to representativeness of data
    - ✧ Compliance with ISO/IEC 27001 (Information security, cybersecurity and privacy protection - Information security management systems - Requirements)
    - ✧ Evaluation based on ISO/IEC 25012 (Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model)
- During the training
  - ➢ Regularization with an additional penalty term for fairness
    - ✧ Using optimization methods with constraints on fairness
  - ➢ RLHF (Reinforcement Learning from Human Feedback)

[62] A feature quantity is a numerical representation of a characteristic of data and is used in machine learning. For example, height and weight correspond to human characteristics. These numerical values are fed to algorithms and used for model learning and prediction.

        - ✧ Learning process to reflect human value criteria and preferences in the output of AI models
- After the training
    - ➢ Implementation of monitoring of data, training process, and results
        - ✧ Considering restructuring of the dataset, including human adjustments to the algorithm as needed and periodic review of the quality and quantity of the dataset to be trained
    - ➢ Proper storage of data and implementation of access controls
        - ✧ Data encryption and secure storage
        - ✧ Compliance with ISO/IEC 27002 (Information security, cybersecurity and privacy protection – Information security controls) for data storage and access

[Ref.]

- Digital Agency "Data Quality Guidebook (beta version)" (June 2021)
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- AI Product Quality Assurance Consortium "AI Product Quality Assurance Guidelines" (June 2023)
- Personal Data +α Study Group "Final Recommendations on Profiling" (April 2022)
- NIST, "AI Risk Management Framework Playbook" (January 2023).

Column 8: Case study on consideration of bias in data during data preprocessing and training[63]

[Use Case Name]
Loan in 7 minutes

[Scope]
A fully automated solution in which AI makes credit decisions in a few minutes based on an analysis of customer behavior and makes optimal loan proposals to customers.

[Scenes in which Data is Handled]
The solution collects all detailed information about the customer, interacting with internal (e.g. transaction data) and external (e.g. credit bureaus) systems, and applies algorithms based on AI and machine learning methods to automatically perform risk estimation and calculate appropriate offers for the customer.

[Implementation Measures]
Utilization of Fairness by Design[64] , a development methodology that considers fairness from the design stage

- Using a participatory design approach that incorporates stakeholder input from the design stage, the AI model can be developed to balance business requirements and fairness by quantifying the weights of attributes such as income, employer, and transaction history, which are the criteria for loan screening, and attributes such as age, gender, and nationality, which are related to fairness. In addition, the method will also incorporate algorithms to reduce cross-biases that appear when attributes such as age, gender, and nationality are combined under certain conditions, as a method to eliminate unacceptable biases in attributes related to fairness, culture, and business practices.

Use of open source software (OSS) technologies by the Intersectional Fairness Project under the Linux Foundation[65] as a countermeasure to potential bias.[66]

- Intersectional Fairness is a bias detection and mitigation technique to address cross-bias caused by combinations of multiple attributes, leveraging existing single attribute bias mitigation methods to ensure fairness in machine learning models with respect to cross-bias.

[63] The following is an example of consideration of bias, etc., in data. This case study is based on a technical report ISO/IEC TR 24030 (ISO/IEC TR 24030) developed by Subcommittee SC 42 (ISO/IEC JTC 1/SC 42) under ISO/IEC JTC1, a technical committee jointly established by ISO, an international standardization organization for which the Japanese Industrial Standards Committee (JISC) represents Japan, and the IEC. 2021), which cites use cases collected in ISO/IEC TR 24030 (2021). (https://www.iso.org/standard/77610.html)
[64] https://pr.fujitsu.com/jp/news/2021/03/31-1.html
[65] The world's largest and most popular open source software project https://www.linuxfoundation.jp/
[66] https://lfaidata.foundation/projects/intersectional-fairness-isf/
https://pr.fujitsu.com/jp/news/2023/09/15.html

[Description in this volume (reprinted)]
During AI development
(D-2) ii. Development that takes into consideration human life, body, property, spirit and environment

- ✧ To prevent harm to the life, body, property, spirit and environment of stakeholders, consider the following ("2) Safety")
  - Performance requirements not only under expected usage conditions in a variety of situations, but also in unanticipated environments
  - Methods to minimize risk (e.g., loss of control of interlocking robots, improper output, etc.) (e.g., guardrail technology)

[Points]

AI developers pay attention to prevent AI systems from causing harm to human life, body, property, mind, and environment by taking countermeasures as necessary, based on the nature and manner of possible damage.

Furthermore, developers are expected to confirm the verification and validity of the AI system in advance in order to assess risks related to its controllability. As a method of risk assessment, experiments may be conducted in a closed space, such as a laboratory or a secure sandbox, before practical application in society.

Developer also pay attention to organize in advance the measures to be taken in the event of harm, if any.

Moreover, in addition to compliance with existing laws, regulations, and guidelines, it is expected to consider the idea of using new technologies to address problems that new technologies may cause.

[Specific Methods]

- Performance requirements for use in unanticipated environments
  - ➢ Implementation of fail-safe features
    - ✧ Design for system migration with safety as a priority in case of failure
  - ➢ Fault-tolerant design
    - ✧ A design policy that allows the system to continue to function and operate by switching to a backup system, etc., even if its components fail or stop
  - ➢ Foolproof design
    - ✧ Design to operate safely even in the event of mishandling
- Minimization of risks (e.g., loss of control of interlocking robots, improper output, etc.)
  - ➢ AI Governance establishment
  - ➢ Guardrail setup
  - ➢ Fallback design
    - ✧ Design policies that allow for partial suspension or reduction of functions when problems occur, such as by running the system on a rule basis or through final human decision making
  - ➢ Consideration and implementation of appropriate mitigation measures to address identified risks and vulnerabilities
  - ➢ Implementation of a phased review process
    - ✧ Preparing detailed checks for AI systems
    - ✧ Conducting reviews based on the checkes in this document throughout the AI lifecycle, including pre-deployment and pre-market
- Adoption of a transparent development strategy
  - ➢ Development of strategies to identify potential risks upstream, such as in development design, and to mitigate risks throughout the development process, for development without compromising safety
- Consideration of measures to be taken in the event of a hazard
  - ➢ Initial response measures
    - ✧ Implemented according to the necessary procedures depending on the context, such as the urgency of the system including the AI in question

  - Restoration through rollback of the system, use of an alternative system, etc.
  - System shutdown
    - ✧ Kill switch
  - Disconnection from the network
  - Confirmation of the nature of the harm
  - Report to relevant parties
  - (In the event of serious damage) Investigation of the causes, analysis, and recommendations by a third-party organization, etc.
- Consideration of new technologies to address risks
  - Development of AI to detect and defend against new cyber attacks
  - Development of AI to remove inappropriate products by AI, etc.

[References]

- Ministry of Internal Affairs and Communications, "Current Status and Issues Concerning Information Distribution in Digital Space" (November 2023)
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- Information-technology Promotion Agency, Japan "SEC journal Vol.10 No.3 Special Issue on Reliability and Security" (September 2014)
- Information-technology Promotion Agency, Japan "White Hacker Study Session for Beginners" (September 2018)
- Personal Data +α Study Group "Final Recommendations on Profiling" (April 2022)
- NIST, "AI Risk Management Framework Playbook" (January 2023).
- The University of Electro-Communications, "Fallback and Recovery Control System of Industrial Control System for Cybersecurity" (October 2017).
- World Economic Forum, "The Presidio Recommendations on Responsible Generative AI" (June 2023).

Column 9: Case study of the use of guardrails to minimize risk

In order to minimize the risk of AI systems, it is expected to consider "guardrails" as a mechanism to control such risks. There are several types of such "guardrails" and they are expected to be utilized according to the matters required at the time of development.

[Examples]

- Topical Rail
  - A method to avoid touching on topics that are not relevant to specific use cases or the intentions of AI business users and non-business users
- Moderation Rail
  - A method to ensure that responses do not contain ethically inappropriate language
- Fact Checking and Hallucination Rail
  - A method to avoid outputting false or illusory answers
- Jailbreaking Rail
  - A method to ensure robustness against malicious attacks

As a specific example of the use of the guardrail method, rinna Corporation offers the Profanity Classification API[67] (an API that detects inappropriate expressions related to discrimination, atrocities, politics, religion, etc. and can be used to monitor SNS, reviews, etc.) for developers. In addition, when they released an image generation model specialized for Japanese and incorporated it into their services, they utilized a safety checking tool called SafetyChecker [68] to check for inappropriate images against generated content[69].

[67] Profanity Classification API
https://developers.rinna.co.jp/api-details#api=profanity-classification-api&operation=profanity-classification-api
[68] SafetyChecker
https://github.com/huggingface/diffusers/blob/main/src/diffusers/pipelines/stable_diffusion/safety_checker.py
[69] Japanese Stable Diffusion (Japanese Stable Diffusion) model card at the time of release: see Safety Module
https://huggingface.co/rinna/japanese-stable-diffusion

[Description in this volume (reprinted)]
During AI development

(D-2) iii. Development that contributes to appropriate utilization

- ✧ To avoid harm caused by provision or use that was not envisioned at the time of development, develop the product by setting the scope of safe use ("2) Safety").
- ✧ Appropriate selection of trained AI models (e.g., whether the license is commercially available, pre-training data, and specifications required for training and execution) when performing post-training on pre-trained AI models ("2) Safety").

[Points]

In developing AI systems, AI developers are expected to cooperate with relevant parties to take preventive measures and follow-up actions (information sharing, suspension/restoration, clarification of causes, measures to prevent recurrence, etc.) according to the nature and manner of damage caused or brought about by accidents, security breaches, privacy violations, etc. that may occur or have occurred when AI is used.

[Specific Methods]

- Guardrail setup
  - ➢ Topical Rail
    - ✧ A method to avoid touching on topics that are not relevant to specific use cases or the intentions of AI users and out-of-business users
  - ➢ Moderation Rail
    - ✧ A method to ensure that responses do not contain ethically inappropriate language
- Alignment of the AI model with objectives
  - ➢ Characteristics of the dataset
    - ✧ Comparing the characteristics of the dataset on which the original model was trained with those on the new task, and consider whether the characteristics trained by the original model are applicable to the new task
  - ➢ Domain of the new task
    - ✧ Confirming that the domain of the new task in which fine tuning is performed matches the domain of the original model. If the domains are different, considering adjustments such as fine tuning only some of the layers
  - ➢ Language consistency
    - ✧ Ensuring that the language of the original model and the new dataset match. If different, considering adjustments such as tokenization methods, vocabulary expansion, etc.

[Ref.]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

[Description in this volume (reprinted)]
During AI development

(D-3) ii. Consideration of biases, etc. contained in algorithms, etc. of AI models

- ✧ Consider even the possibility that bias can be included by each technical component of the AI model (e.g., prompts input by AI users and off-business users, information referenced during inference of the AI model, and external services to be linked) ("3) Fairness").
- ✧ Given that bias cannot be completely eliminated from AI models, development based on a variety of methods rather than a single method should be conducted in parallel ("3) Fairness").

[Points]

AI developers are aware that the training algorithms used in AI may result in biass in the output of AI.

In addition, in order to maintain the fairness of the results of judgments made by AI, human judgment is expected to intervene with regard to whether or not to use the AI judgment or how to use the AI judgment, based on the social context in which AI is used, people's reasonable expectations, and the significance of the judgment on the rights and interests of those who are subject to the judgment using AI.

[Specific Methods]

- Bias detection and monitoring
  - ➢ Attention to prompts entered by AI business users
    - ✧ Reminding AI providers of the need to conclude terms of use with AI business users
  - ➢ Verification of information and external services during inference, etc.
- Review of feature
  - ➢ Clarification of sensitive attributes (personal attributes such as gender and race of the subject that should be excluded from the perspective of fairness) for each business
    - ✧ In clarifying these attributes, considering the grounds enumerated in Article 14(1) of the Constitution and the attributes referred to in international human rights rules
  - ➢ Clarification of the content of fairness to be ensured with respect to sensitive attributes
    - ✧ Collective fairness
      - Removing sensitive attributes and make predictions based only on non-sensitive attributes (unawareness)
      - Ensuring the same prediction results across groups with different values of sensitive attributes (demographic parity)
      - Adjusting the ratio of the error of the predicted result to the actual result independent of the value of the sensitive attribute (equalized odds)
    - ✧ Individual fairness
      - Giving the same predicted result to each individual with equal values of non-sensitive attributes
      - Giving similar prediction results to individuals with similar attribute values (fairness through awareness)
- Use of bias-aware models in machine learning models
  - ➢ Use of Inverse Probability Weighting (IPW)
    - ✧ A method to ensure equality by weighting the collected data by groups, etc.
- Achievement of fairness in machine learning systems (from qualitative to quantitative approaches)
  - ➢ AI developers consider realizing the fairness risks analyzed by the AI provider through quantitative fairness metrics such as "equality of outcome" from the implementation stage, if necessary

- Human judgment intervention based on social context and people's reasonable expectations
  - When statistical future predictions are difficult to make (due to high uncertainty)
  - When a convincing reasons are required for a decisions (judgments), such as when there is a significant impact on a specific individual or group
  - When discrimination based on race, creed, or gender is expected due to social bias against minorities in the training data, etc.

[Ref.]
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

[Description in this volume (reprinted)]
During AI development

D-5) i. Implement mechanisms for security measures

- ✧ Throughout the process of developing AI systems, take appropriate security measures in light of the characteristics of the technology to be employed (security by design) ("5) Ensuring Security").

[Points]

It is expected to pay attention to the security of AI and take reasonable measures to ensure the confidentiality, integrity, and availability of AI systems, in light of the technical level at that time. In addition, measures to be taken in the event of a security breach are expected to be organized in advance, taking into account the applications and characteristics of the AI in question, the magnitude of the impact of the breach, and other factors.

Ensure the security of the developed system by considering security from the early stage of the development process, referring to Security by Design, etc. defined by the National center of Incident readiness and Strategy for Cybersecurity (NISC). Security is to be ensured by considering security from the early stages of the development process. Conventional methods of adding security functions as an afterthought or implementing security tools just before shipment may cause frequent rework and result in high development costs. Implementing security measures at an early stages of development reduces rework and costs, and leads to the creation and provision of systems and software with good maintainability.

Machine learning systems, including LLM, require further analysis methods and improved control measures in addition to conventional information systems due to the nature of the assets (training data sets, training models, parameters, etc.), stakeholders (training model providers, etc.), and probabilistic outputs. Therefore, it is important to develop and apply security analysis methods and control measures based on the technical characteristics of machine learning.

[Specific Methods]

- Security by Design
  - ➢ Examples of security measures
    - ✧ Threat Assessment
      - Clarify the threats and possible attacks that software faces, and clarify what to protect software from
    - ✧ Security Requirements
      - Defines the secure behavior of the software itself. Types of requirements include requirements for system functionality, usability, maintainability, performance, etc. Security requirements are the requirements related to security among the system requirements, and define the objectives necessary for safe operation of the system. Describe security requirements as a part of the system requirement definition document or as a security requirement definition document
    - ✧ Security Architecture
      - Use the recommended architecture by the platform provider for the AI system, customizing it, rather than developing your own architecture
    - ✧ Software Bill of Materials (SBOM)
      - SBOM creation is incorporated to facilitate visualization and configuration management of the software embedded in the product.
    - ✧ Responsible use of open source software
      - Responsible use of open source software involves screening open source packages, facilitating code contributions for dependencies, and cooperating to keep critical components developed and maintained
- Enhanced security measures
  - ➢ Risk assessment
    - ✧ Conduct risk assessments for information security to identify and prioritize risks

- ISO/IEC 27001: Information security, cybersecurity and privacy protection - Information security management systems - requirements
- SP800-30: Guide for Conducting Risk Assessments

- Access control and authentication
  - Grant minimum necessary access rights and employ strict authentication measures for developers and administrators to access the system
    - ISO/IEC 27001: Information security, cybersecurity and privacy protection - Information security management systems - Requirements
    - SP800-53: Security and Privacy Controls for Information Systems and Organizations
  - Establish a robust internal threat detection program for intellectual property and trade secret content that is important to each actor
- Awareness and training
  - Conduct awareness education and training to ensure fulfillment of cybersecurity obligations and responsibilities in accordance with relevant policies, procedures, contracts, etc.
    - ISO/IEC 27002: Information security, cybersecurity and privacy protection – Information security controls
    - SP800-50: Building an Information Technology Security Awareness and Training Programs
- Ensuring data security
  - Use encryption when transferring data and apply security protocols when storing and processing data
    - ISO/IEC 27001: Information security, cybersecurity and privacy protection - Information security management systems - Requirements
    - SP800-53: Security and Privacy Controls for Information Systems and Organizations
- Processes and procedures for protecting information
  - Maintain security policies, processes, and procedures, and used them to manage the protection of information systems and assets
    - ISO/IEC 27001: Information security, cybersecurity and privacy protection - Information security management systems - Requirements
    - SP800-37: Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy
- Attention to bugs in open source applications, etc.
  - Promptly update information on bugs, etc. contained in open source
    - ISO/IEC 27009: Information security, cybersecurity and privacy protection - Sector-specific application of ISO/IEC 27001 - Requirements
    - SP800-161: Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations
- Maintenance
  - Perform maintenance work and record maintenance work using approved and controlled tools
    - ISO/IEC 27001: Information security, cyber security and privacy protection - Information security management systems - Requirements
    - SP800-40: Guide to Enterprise Patch Management Planning: Preventive Maintenance for Technology
- Monitoring and incident response
  - Establish a monitoring system and implement an incident response process when anomalies are detected
  - Appropriately document incidents as they occur and consider mitigating identified risks and vulnerabilities
    - ISO/IEC 27001: Information security, cybersecurity and privacy protection - Information security management systems - Requirements
    - SP800-61: Computer Security Incident Handling Guide

- See "Table 4: Examples of each attack causing damage to machine learning systems" for examples of attack techniques

Table 4: Examples of Damage and Threats to Machine Learning Utilization Systems[70]

Damage			Threats that cause damage: Threats Specific to Machine Learning	Threats that cause damage: Other Threats
Integrity or availability violations	System malfunction	Due to unintended behavior of machine learning elements	data poisoning attack	Conventional attacks on software and hardware implementing machine learning elements
			Model Poisoning Attack	
			Pollution Model Abuse	
			evasive attack	
		Due to other factors		Conventional attacks against the system
	waste of computing resources	By machine learning elements	Data poisoning attack (resource exhaustion type)	Conventional attacks on software and hardware implementing machine learning elements
			Model Poisoning Attack (resource depletion type)	
			Pollution Model Abuse	
			sponge attack	
		Due to other factors		Conventional attacks against the system
Breach of Confidentiality	Leakage of information about AI models		model extraction attack	Conventional attacks to steal AI models
	Leakage of sensitive information contained in training data		Information leak attack on training data	Conventional attacks that steal data
			Data Poisoning Attacks (Information Embedded)	
	Leakage of other confidential information		Model Poisoning Attack (Information Embedded)	

[Ref.]

- Ministry of Economy, Trade and Industry, "Casebook on OSS Utilization and Management Methods to Ensure Its Security" (April 2021)
- Ministry of Economy, Trade and Industry, "Guidance on the Introduction of SBOM for Software Management" (July 2023)
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- Information-technology Promotion Agency, Japan "Security by Design Installation Guide" (August 2022)
- NCSC, "Guidelines for secure AI system development" (November 2023).
- NIST, "CYBERSECURITY FRAMEWORK" (April 2018).
- ISO/IEC27000 Series
- NIST, SP800 Series

[70] Adapted from "Machine Learning Quality Management Guidelines, Fourth Edition" (December 2023), National Institute of Advanced Industrial Science and Technology (AIST)

[Description in this volume (reprinted)]
During AI development

(D-6) i. Ensure verifiability

- ✧ Considering the characteristics that the prediction performance and output quality of AI may fluctuate significantly after the start of utilization and may not reach the expected accuracy, maintain and improve the quality of AI while preserving work records for post-verification ("2) Safety" and "6) Transparency").

[Points]

AI developers are expected to record and store logs during development to ensure verifiability of AI input/output, etc., and to develop AI systems in such a way that AI providers and others can obtain logs of input/output.

AI developers are expected to design and develop AI systems in a manner that ensures transparency so that AI providers can understand and appropriately offer AI systems to AI business users.

[Specific Methods]

- Record and store logs
  - ➢ Specifically, record and store the following logs
    - ✧ What data was used during AI development? etc.
  - ➢ When considering the necessary "logs," also refer to the "documents" and "records" required by the management system or contract to which your organization is certified, and record and store the appropriate logs. Specifically, consider the following
    - ✧ Purpose of logging and storage
    - ✧ Accuracy of logging
    - ✧ Frequency of logging and recording
    - ✧ Time of logging, duration of storage, and capacity of storage location
    - ✧ Protection of logs
      - Ensure confidentiality, integrity, availability, etc.
    - ✧ Scope of logs to be disclosed, etc.
- Consider methods to increase accountability and interpretability. Note that in considering the following, there may be concerns about trade-offs with development
  - ➢ Use of simple models
    - ✧ Select the simplest possible model for the requirements, if possible
      - Logistic regression, decision tree, etc.
  - ➢ Local explanatory methods
    - ✧ Use of local explanatory techniques in explaining model predictions
    - ✧ A method that describes the behavior of a model for a specific data point, most notably LIME (Local Interpretable Model-agnostic Explanations), etc.
  - ➢ SHAP values (SHapley Additive exPlanations)
    - ✧ Evaluates how much each feature contributes to the model's predictions based on game theory, making it easier to understand the relative impact of each feature
  - ➢ Feature contribution visualization
    - ✧ Use a method to visualize features that model considers important
      - This includes feature importance plots, partial dependence plots, etc.
  - ➢ Analysis of model internals
    - ✧ Employs a method for detailed analysis of the model's internal structure and behavior
    - ✧ Frameworks such as TensorFlow and PyTorch also allow visualization of output and gradients in the middle layer of the model
  - ➢ Selection of model architecture
    - ✧ Choose model architectures carefully to emphasize interpretability
  - ➢ Consideration of stakeholder participatory methods

        - ✧ Incorporate feedback from stakeholders (e.g., AI providers and AI business users) and knowledge from domain experts
    - ➢ Introduction of watermarks that clearly indicate the use of AI, where technically feasible
        - ✧ Consider labeling, disclaimers, and other mechanisms to ensure that AI business users and non-business users are aware of their interactions with AI systems
- Analysis of trends in AI output based on multiple input/output combinations to the AI
    - ➢ For example, observing changes in output when the input pattern is varied gradually, etc.

[Ref.]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- AI Product Quality Assurance Consortium "AI Product Quality Assurance Guidelines" (June 2023)
- ISO, "ISO/IEC 23894:2023 (Information technology-Artificial intelligence-Guidance on risk management)" (February 2023)
- The White House, "Blueprint for an AI Bill of Rights (Notice and Explanation)" (March 2023).
- World Economic Forum, "The Presidio Recommendations on Responsible Generative AI" (June 2023).

[Description in this volume (reprinted)]
After AI development

(D-5) ii. Attention to the latest trends

- ✧ New attack methods against AI systems are emerging every day. To address these risks, check points to be considered in each development process ("5) Ensuring Security").

[Points]

AI developers are expected to keep abreast of the latest developments in order to gain technical insight and implement more advanced and sustainable AI development.

Through this, AI developers are expected to work with AI providers to ensure that AI systems are utilized as intended, with appropriate timing commensurate with the degree of risk, and to monitor vulnerabilities, incidents, new risks, and exploits after deployment and take appropriate steps to address them.

[Specific Methods]

- Stay up-to-date on the latest developments through
  - Papers from international conferences, arXiv, etc.
  - JVN iPedia Vulnerability Countermeasure Information Database
  - SNS and other developer communities
  - Refer to open source projects
  - Press, etc.

[Ref.]

- Ministry of Internal Affairs and Communications "AI Security Information Dissemination Portal[71]
- Information-technology Promotion Agency, Japan "JVN iPedia Vulnerability Countermeasure Information Database" [72]
- Cornell University, "arXiv".[73]

[71] Ministry of Internal Affairs and Communications, "AI Security Information Dissemination Portal," https://www.mbsd.jp/aisec_portal/
[72] Information-technology Promotion Agency, Japan "JVN iPedia Vulnerability Countermeasure Information Database", https://jvndb.jvn.jp/index.html
[73] Cornell University, "arXiv",https://arxiv.org

[Description in this volume (reprinted)]
After AI development

(D-6) ii. Provide information to relevant stakeholders

- ✧ Ensure that you can explain your AI system to relevant stakeholders (including through your AI provider) in a timely and appropriate manner, for example, with respect to the following ((6) Transparency)
  - Possibility of changes in output or programs due to learning, etc. of AI systems ("1) Human-centric")
  - Information on safety, including technical characteristics of the AI system, mechanisms for ensuring safety, foreseeable risks that may arise as a result of its use, and mitigation measures ("2) Safety")
  - Scope of use intended by the AI developer to avoid harm due to provision or use not envisioned at the time of development ("2) Safety")
  - Information on the operational status of the AI system, causes of and responses to defects ("2) Safety")
  - Information on what, if any, updates have been made to the AI system and why ("2) Safety")
  - Collection policy of data to be trained by the AI model, its learning method and implementation system, etc. ("3) Fairness", "4) Privacy protection", and "5) Ensuring security")

[Points]

AI developers are expected to explain the status of compliance with Common Guiding Principles for the AI systems they develop (including through AI providers), for the purpose of gaining stakeholders‘ understanding and reassurance, as well as to present evidence for the AI's operation for the purpose.

This does not assume disclosure of algorithms or source codes themselves, but is expected to be conducted within a reasonable range in light of the characteristics and applications of the adopted technology, while taking privacy and trade secrets into consideration.

[Specific Methods]

- Indication that AI is being used
- Develop and clarify AI policy on ethics
  - ➢ Publicize ethical principles and policies and make clear the commitment of AI developers to a code of ethics
    - ✧ Includes disclosure of policies regarding personal data, user prompts, AI system output, privacy, etc., to the extent reasonable
  - ➢ For details, see Appendix 2. Goal 2-1 [AI Governance Goal Setting] of Action Objectives.
- Dialogue with Stakeholders
  - ➢ Conducting dialogue with stakeholders while providing information on ethical initiatives and transparency through websites, etc.
    - ✧ For details, see Appendix 2. Action Objective 5-2 [Consideration of External Stakeholders' Opinions].
  - ➢ Consider a mechanism to encourage reporting to AI developers when relevant stakeholders discover problems or vulnerabilities after implementation.
    - ✧ For example, after the introduction of the system, incentives such as a reward system for incident reporting should be provided to promote vulnerability discovery through relevant stakeholders, etc.
    - ✧ For details, see Appendix 2. Action Objective 3-4-2 [Preliminary Review of Response to Incidents/Conflicts].

[Ref.]

- EU, "Ethics Guidelines for Trustworthy AI" (April 2019).

- ISO, "ISO/IEC 23894:2023 (Information technology-Artificial intelligence-Guidance on risk management)" (February 2023)

[Description in this volume (reprinted)]
After AI development

(D-7) i. Explanation to AI providers of the status of compliance with the "Common Guiding Principles

- ✧ Provide AI providers with information and explanations on the characteristics of AI, including the possibility of significant fluctuations in forecasting performance and output quality after the start of utilization, and the risks that may arise as a result of such fluctuations. Specifically, the following items are to be made known to the public ("7) Accountability").
  - Addressing possible bias in each technical component of the AI model (training data, training process of the AI model, prompts assumed to be input by AI users and non-business users, information referenced during inference of the AI model, external services to be linked, etc.) ("3) Fairness").

[Points]

AI developers are expected to provide AI providers with meaningful and useful information and explanations appropriate to the social context and the size of the risk, as well as to disclose reports and other information in an understandable format regarding the development and technical evaluation as much as possible.

This does not assume disclosure of algorithms or source codes themselves, but is expected to be conducted within a reasonable range in light of the characteristics and applications of the adopted technology, while taking privacy and trade secrets into consideration.

[Specific Methods]

- Provide an explanation of the status of the response within the scope that does not violate trade secrets, etc., while implementing the followings when trade-offs arise
  - ➢ Evaluating whether/to what extent non-disclosure is acceptable from the perspective of transparency, ethics, etc.
  - ➢ Documenting the decision-making process
  - ➢ Holding decision-makers responsible for their decisions
  - ➢ Providing appropriate and ongoing oversight of the decisions

[Ref.]

- EU, "Ethics Guidelines for Trustworthy AI " (April 2019).
- ISO, "ISO/IEC 23894:2023 (Information technology-Artificial intelligence-Guidance on risk management)" (February 2023)

[Description in this volume (reprinted)]
After AI development

(D-7) ii. Documentation of development-related information

- ✧ To improve traceability and transparency, document the development process of AI systems, data collection and labeling that influence decision making, algorithms used, etc., in a manner that allows third-party verification whenever possible ("7) Accountability")

(Note: This does not mean that everything documented here will be disclosed.

[Points]

AI developers are expected to maintain/retain the AI development process and reported incidents, etc., with appropriate documentation, while working with stakeholders as necessary, to ensure third-party verifiability and to mitigate identified risks and vulnerabilities.

[Specific Methods]

- Documentation
  - ➢ Documentation of the development process
    - ✧ Provide a reasonable explanation of how decisions were made, beginning with the source of the data, and keep records such as transparency reports to ensure traceability
    - ✧ In implementing the above, keep in mind that in the event of an unforeseen incident in an AI system, all parties in the AI value chain may be in a position in the future to be asked to provide some explanation
  - ➢ Documentation of reported incidents
    - ✧ Consider documenting incidents appropriately to mitigate identified risks and vulnerabilities
- Documentation methods
  - ➢ These documents are updated regularly
  - ➢ The format and medium of documentation are chosen by each actor. It does not necessarily have to be paper-based
  - ➢ Be made available to stakeholders depending on the context of use

[Ref.]

- ISO, "ISO/IEC 23894:2023 (Information technology-Artificial intelligence-Guidance on risk management)" (February 2023)

[Description in this volume (reprinted)]
After AI development

(D-10) i. Contribution to the creation of innovation opportunities

- ✧ To the extent possible, the following items are expected to contribute to the creation of opportunities for innovation ("10) Innovation")
  - Conduct research and development on quality and reliability of AI, methodology for development, etc.
  - Contribute to maintaining sustainable economic growth and presenting solutions to social issues
  - Internationalize and diversify and collaborate with industry, academia, and government by referencing trends in international discussions such as DFFT, participating in AI developer communities and conferences, and other initiatives.
  - Provide information on AI to society as a whole.

[Points]

AI developers are expected by society in particular to drive innovation, as they can directly design and make changes to AI models and have a high impact on the output of AI in overall AI systems/services

[Specific Methods]

- Develop and promote shared standards, tools, mechanisms, and best practices to ensure the safety, security, and reliability of AI systems, and establish mechanisms to adopt them as needed
  - ➢ Share best practices among organizations to improve safety and ensure security
  - ➢ Collaborate with stakeholders including industry, academia, government agencies, and non-profit organizations

B. Explanation of "Common Guiding Principles" in "Part 2"

This section describes specific techniques that are not mentioned in "Part 3: Matters Relating to AI Developers" in this volume, but are of particular importance to AI developers among the "Common Guidelines" in "Part 2" in this volume.

The AI developer shall take necessary measures such as providing necessary information when requested by the AI provider or AI user.

[Contents of this volume (reprinted)] * Only the columns are excerpted.

1) Human-centered

Each entity should ensure that the development, provision, and use of AI systems and services do not violate at least constitutionally guaranteed or internationally recognized human rights as a foundation from which all matters to be addressed, including each of the items described below, are derived. It is also important to act in such a way that AI expands people's capabilities and enables the pursuit of diverse happiness (well-being) of diverse people.

- Related to "(1) Human Dignity and Individual Autonomy
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - Respect human dignity and individual autonomy, taking into account the social context in which AI will be used
  - In particular, when linking AI with the human brain and body, refer to discussions of bioethics in other countries and research institutions, while taking into account information on peripheral technologies.
  - When profiling using AI in areas that may have a significant impact on the rights and interests of individuals, respect the dignity of individuals, maintain the accuracy of the output, understand the limitations of AI in terms of prediction, recommendation, or judgment, and carefully consider any possible disadvantages before using AI. Do not use it for inappropriate purposes.

  [Specific Methods]
  - Establish a director in charge of AI ethics and an internal organization on AI governance
  - Examples of bioethics discussions in other countries and institutions that can be referenced during AI development include
    - Reports issued by international organizations such as the United Nations (UN) and the World Health Organization (WHO), etc.
    - Research papers published by universities and other academic institutions
  - When profiling using AI in areas that may have a significant impact on the rights and interests of individuals, it is particularly useful to pay attention to the following points
    - Minimize any potential bias in the data or algorithms used for profiling to obtain fair and equal results
      - Monitoring of data and algorithm outputs
      - Ensure that the individual concerned has the opportunity to receive human judgment as well as AI judgment, etc.
    - If personal information is included in the data used for profiling, to handle such personal information appropriately
      - Establishment of a data clean room
      - Implementation of privacy-preserving machine learning, etc.
    - Ensure that the AI systems developed function properly and adequately manage potential risks to individuals

- Related to "(2) Consideration of decision-making and emotional manipulation by AI
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - We will not develop, provide, or use AI systems or services for the purpose of, or on the premise of, improperly manipulating human decision-making, cognition, or other emotions.
  - Take necessary measures against the risk of over-reliance on AI, such as automation bias[74] , in the development, provision, and use of AI systems and services.
  - Be wary of the use of AI that encourages information and value tilting, as typified by filter bubbles, and unwillingly limits the choices that humans, including AI users, should have available to them.
  - Handle AI outputs with caution, especially when they may be relevant to procedures that have a significant impact on society, such as elections and community decision-making.

  [Specific Methods]
  - As a measure to address the risk of over-reliance on AI to automation bias, etc., it is useful to request AI providers to alert AI users and off-business users
  - For example, serendipity can be a useful way to deal with filter bubbles.
    - Specifically, use of various information sources, review of algorithms, etc.
  - In cases that can be related to procedures that have a significant impact on society, such as elections and community decision-making, for example, it is useful to evaluate AI systems from an ethical rather than a technical perspective, for example, by having humans implement the final decision.

- Related to "(3) Countermeasures against false information, etc.
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - We recognize that the risk of AI-generated disinformation, misinformation, and biased information destabilizing and confusing society is increasing, and we must take necessary countermeasures.

  [Specific Methods]
  - For example, in addition to guardrails such as fact-checking (a mechanism to investigate the truth and accuracy of factual statements and report the results of verification), it is useful to indicate that the product is generated by a generation AI, etc.

- Related to (4) Ensuring Diversity and Inclusion
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - In addition to ensuring fairness, care will be taken to facilitate the use of AI by the socially vulnerable so that more people can enjoy the benefits of AI without creating so-called "information and technology weaklings".

  [Specific Methods]
  - For example, universal design, ensuring accessibility, education and follow-up with relevant stakeholders, etc. are useful

- Related to (5) User Support
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - To the extent reasonably possible, provide information on the functionality of the AI system/service and its surrounding technology, with the ability to provide timely and appropriate information for determining the opportunity for selection, in a state of availability

[74] It refers to the phenomenon of excessive trust and reliance on automated systems and technologies in human judgment and decision-making.

[Specific Methods]
- Explanation of information on data handling
  - How to utilize the data input into the AI model created by the AI developer for additional learning, etc.
  - Information on the source and processing of the data used in the study
- Ensure transparency of algorithms and AI models
  - Disclosure of algorithmic logic, if possible
  - Examples of inputs and outputs
- Notification of changes and updates

- Related to (6) Ensuring sustainability
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - In the development, provision, and use of AI systems and services, the impact on the global environment is also considered throughout the lifecycle.

  [Specific Methods]
  - Adoption of lightweight AI model
    - Increased energy efficiency by using lightweight, resource-efficient AI models instead of large, highly accurate AI models, in line with your AI requirements
  - Optimization of AI model size
    - AI model design and algorithm development with awareness of efficient use of computational resources and minimized energy consumption.
  - Effective use of data
    - Improve data quality, eliminate redundancy, and avoid unnecessary data acquisition

[Ref.]
- Ministry of Internal Affairs and Communications, "Global and Japanese Situations and Issues Surrounding Fact Checking" (May 2019)
- EU, "Ethics Guidelines for Trustworthy AI" (April 2019).
- OIS Research Conference, "AI and Citizen Science for Serendipity" (May 2022)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(4) Privacy Protection Considerations

It is important for each entity to respect and protect privacy in the development, provision, and use of AI systems and services, depending on their importance. In doing so, they should comply with relevant laws and regulations.

- Related to "(1) Protection of privacy in AI systems and services in general
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - Take actions appropriate to the importance of stakeholders' privacy to ensure that it is respected and protected, taking into account the social context and people's reasonable expectations, by complying with the Personal Information Protection Law and other relevant laws and regulations, and by developing and publishing a privacy policy for each entity.

  [Specific Methods]
  - Enhanced security measures to protect privacy (see "Table 5. Examples of machine learning-specific threats, attack interfaces, attack execution phases, attackers, and attack methods" for more information on machine learning-specific attack methods)
    - Implement appropriate encryption methods and access control mechanisms
    - Testing and fine-tuning to ensure that personal information is not divulged
    - For those of particular importance, consider introducing privacy-preserving machine learning, secure machine learning, etc.
  - Consider implementing data management and restriction functions to control access to data
    - Introduction of authorization for data access
    - Setting up a data management organization
      - Establishment of CDO (Chief Data Officer)
      - Designation of Privacy Officer
      - Dedicating resources to privacy initiatives
        - Staffing, human resource development, etc.
    - Establishment and dissemination of data operation rules
  - Conducting Privacy Assessments
    - Privacy Impact Assessment (PIA)
      - Visualize and organize information collected and processed by AI systems, information flows, and stakeholders
      - Identify privacy risks to AI systems
      - Determine the impact and likelihood of occurrence of each risk and assess the magnitude of risk
      - Determine the direction of risk response (reduction, avoidance, acceptance, or transfer) depending on the magnitude of the risk, and develop a response plan
    - Quality management implementation items (see "Table 6. Summary of Quality Management Implementation Items")
  - Acquisition of ISO standards related to the handling of personal information
    - ISO/IEC 27001."
      - International standard for information security management systems (ISMS), focusing on the maintenance and management of information security
    - ISO/IEC 27701."
      - Describes extended requirements for Personal Information Management Systems (PIMS) based on ISO/IEC 27001
      - Standards focused on privacy protection and can be used by AI developers to ensure proper management of personal information
    - ISO/IEC 29100."
      - International standard for privacy, providing basic principles and requirements for the protection of personal information

- ✧ ISO/IEC 27018."
  - International Standard for the Protection of Personal Information in Cloud Services
  - Can be used by AI developers providing cloud services to ensure appropriate handling of personal information in the cloud environment

[Ref.]

- Ministry of Economy, Trade and Industry, "Information Security Management Standards (2016 Revision)" (March 2016)
- Information-technology Promotion Agency, Japan "SEC journal No. 45, Preface" (July 2016)
- Information-technology Promotion Agency, Japan "How to Create a Secure Website" (March 2021)
- ISO, "Guidelines for privacy impact assessment."
- Northwestern University, "Secure Machine Learning over Relational Data" (September 2021)
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

Table 5. examples of machine learning-specific threats, attack interfaces, attack execution phases, attackers, and attack methods[75]

threat	Attack interface assets	Execution phase of the attack	Examples of attackers	Typical examples of attack techniques
data poisoning attack	Source of training data	During collection and processing of the training data set	External Attackers	Modification of training data collection source
	training data set	During collection and processing of training data sets During system development	Data providers System Developer External attacker	Modification of the training data set
Model Poisoning Attack	pre-learning model	When learning and providing pre-learning models During system development	AI Model Providers System Developers External attacker	Backdoor to pre-training model
	learning mechanism	During system development	System Developers External Attacker	Malicious training programs
	Trained AI Model	During system development During system operation		Modification of AI model
Model Pollution Abuse	Source of operational input data Operational input data system	During system operation	System Users System Operators	Operational input to exploit backdoors (Observation of output information, etc. during operation (to steal information embedded in the model)
model extraction attack	Source of operational input data Operational input data system	During system operation	System Users System Operators	Input data to the system during operation Observation of output information, etc. during operation
Evasive Attacks Sponge attack	Trained AI Model	When obtaining a trained AI model	system operator	Malicious data input to the system during operation Observation of output information, etc. during operation

[75] Adapted from "Machine Learning Quality Management Guidelines, Fourth Edition" (December 2023), National Institute of Advanced Industrial Science and Technology (AIST)

	Source of operational input data Operational input data	During system operation	Input data provider at the time of operation System operator	Modification of data to the system during operation
	system		System Users System Operators	Malicious data input to the system during operation Observation of output information, etc. during operation
Information leak attack on training data	pre-learning model	After obtaining the pre-training model	Model User (System developers)	Observation of input/output and internal information of the obtained AI model in operation
	Trained AI Model	After obtaining a trained AI model	system operator	
	Source of operational input data Operational input data	During system operation	Input data provider at the time of operation System operator	Modification of input data during operation
	system		System Users System Operators	Malicious data input to the system during operation Observation of output information, etc. during operation

Table 6. Summary of Quality Management Implementation Items

Preliminary analysis (primarily AI providers)	essential protected data		Handling of deliverables
	Compliance with applicable laws Identification of personal information requiring special consideration		Determination of reusable deliverables Confirmation of Consent Arrangements
method study (mainly AI developers)	Pre Stage	In Stage	Post Stage
	Learning data quality Protective processing Data distribution (outliers)	Generalization performance PPML (differential privacy)	Setting Safeguards
	trade-off analysis		
	Accuracy vs. fairness of judgment results Data protection measures VS usefulness		

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(6) Transparency Considerations

In developing, providing, and using AI systems and services, it is important for each entity to provide information to stakeholders to the extent necessary and technically feasible and reasonable, while ensuring the verifiability of AI systems and services, taking into account the social context in which AI systems and services are used. The following is a brief overview of the key points of this process.

- Related to "(iii) Reasonable and sincere response."
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - (2) Provision of information to relevant stakeholders" (as described in this Part) does not assume disclosure of algorithms or source code, but is implemented to the extent that it is socially reasonable in light of the characteristics and applications of the technology to be employed, while respecting privacy and trade secrets.
  - When using publicly available technologies, comply with the respective rules and regulations.
  - Consider social impact when open sourcing developed AI systems.

  [Specific Methods]
  - To protect privacy and trade secrets, it is useful, for example, to prepare explanatory documents for non-technical personnel
  - Providing information once is not the end of the process. It is useful to have a series of dialogues with stakeholders to the extent possible in light of the characteristics and applications of the technology to be adopted. It is also important to proactively design and maintain communication design for this purpose, as well as design and development.
  - Check and comply with licenses when using publicly available technologies and libraries
    - Particular attention should be paid to non-commercial licenses, etc.
  - When open-sourcing developed AI systems, it is useful to identify the possible social impact of disclosure and risks through interviews with relevant stakeholders, etc., and to respond to them.

- Related to "(iv) Improvement of accountability and interpretability to relevant stakeholders
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - Analyze and understand what kind of explanations are required and take necessary actions in order to obtain the relevant stakeholders' sense of conviction and reassurance, and to present evidence for the AI's operation for this purpose.
    - AI providers: share with AI developers what explanations will be required
    - AI users: share with AI developers and AI providers what explanations will be required

  [Specific Methods]
  - For example, it is useful to prepare explanatory materials for non-technical users on the operating principles and decision-making process of AI.
    - For other points to keep in mind when explaining, please refer to Appendix 2, Action Objective 4-1 [Ensure accountability of AI management system operation status].

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(7) Accountability Considerations

In the development, provision, and use of AI systems and services, each entity should provide stakeholders with information on how to ensure traceability and how the "common guidelines" are being addressed, to a reasonable extent, based on the role of each entity and the degree of risk posed by the AI systems and services it develops, provides, and uses. It is important to fulfill accountability to a reasonable extent.

[excerpts from relevant descriptions (restatement of descriptions in this volume), and specific methods].

- Related to "(1) Improvement of traceability
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - Ensure that the source of data and decisions made during the development, provision, and use of AI are traceable and retraceable to the extent technically possible and reasonable

  [Specific Methods]
  - Data lineage (construction of a comingled mechanism)
    - Knowing where the data came from, how it was collected and managed, and how it moved within each entity over time
    - Such data includes the identifier of the service or AI model that created the content, but need not include user information
  - Indication that the content is AI-generated (content authentication)
  - Version control of AI models
  - Logging of the training process
  - Backtracking and update history tracking

- Related to "(3) Clarification of Responsible Persons
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - Establish a person responsible for accountability in each entity

  [Specific Methods]
  - When establishing responsible persons, it is useful to clearly define roles and responsibilities
  - Collaborate with AI providers, as necessary, in developing policies to manage risks associated with the use of AI systems and to ensure their safety
  - As for publicizing the above policies, it is useful to use the websites of each entity so that stakeholders can easily access them.

- Related to "(iv) Distribution of responsibility among related parties
  [Relevant descriptions (reiteration of descriptions in this volume)].
  - Clarify the responsibilities among related parties through contracts or social commitments (voluntary commitments) between each entity, including non-working users.
  -
  [Specific Methods]
  - For clarification of responsibilities through contracts, it is useful to refer to "Appendix 6. Key Considerations when Referring to the "Guidelines for Contracts for the Use of AI and Data"" as necessary.
  - Social commitments could take the form of, for example, the formulation of a code of ethics in cooperation with industry associations, etc.

- Related to "(5) Specific Responses to Stakeholders
  [Relevant descriptions (reiteration of descriptions in this volume)].

- Develop and publicize, as necessary, AI governance policies, privacy policies, and other policies for each entity to manage risks associated with the use of AI systems and services and to ensure safety (including social responsibilities such as sharing the vision and disseminating and providing information to society and the general public)
- Provide opportunities for stakeholders to point out errors in AI output, etc., as necessary, and conduct objective monitoring
- If a situation arises that is detrimental to stakeholders' interests, develop a policy on how to respond and steadily implement it, reporting progress to stakeholders on a regular basis as necessary.

[Specific Methods]

- Provide opportunities for feedback from stakeholders, such as a website or contact point for inquiries
  - For details, see Appendix 2, Action Objective 5-2 [Consideration of External Stakeholders' Opinions].
- To the extent possible, regularly publish monitoring results of AI systems
- It is useful to develop a crisis management response plan, etc., in preparation for situations that may harm stakeholders' interests.
  - For details, see Appendix 2, Action Objective 3-4-2 [Preliminary review of response to incidents/conflicts

[Ref.]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- World Economic Forum, "The Presidio Recommendations on Responsible Generative AI" (June 2023).

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(8) Notes on education and literacy

Each entity is expected to provide the necessary education to ensure that those involved in AI within the entity have the knowledge, literacy, and sense of ethics to have a correct understanding of AI and socially correct use of AI. Each entity is also expected to educate its stakeholders, taking into account the characteristics of AI such as complexity and misinformation, as well as the possibility of intentional misuse.

[Relevant descriptions (reiteration of descriptions in this volume)].

- Take necessary steps to ensure that those involved in AI within each entity have a sufficient level of AI literacy in their involvement
- As the segregation of AI and human work is expected to change with the expanded use of generative AI, education, reskilling, etc. will be considered to enable new ways of working.
- Provide educational opportunities to help people from all walks of life better understand the benefits gained from AI and increase their resilience to risk, taking into account the generation gap.
- Provide necessary follow-up to stakeholders to ensure education and literacy, as needed, to enhance the overall safety of AI systems and services.

[Specific Methods]

- Education for AI developers
  - Fostering a mindset and culture that is willing to change and continue to learn about the latest attack methods, etc.
  - Promote collaboration throughout the value chain and understand the trade-offs of collaboration
  - Appealing to the growing need for social responsibility, etc.
- Education for AI providers, AI users and off-business users, etc.
  - Educate AI users and off-business users about the appropriate use of AI systems and potential risks
  - Disseminate information on AI systems being developed by AI developers to increase literacy about their appropriate use, benefits to be gained, potential risks, and how to respond to risks.

[Ref.]

- Tentative Discussion Paper on AI, Cabinet Office (May 2023)
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- NIST, "AI Risk Management Framework Playbook" (January 2023).

C. Items to be observed in the development of advanced AI systems

For AI developers developing advanced AI systems, including state-of-the-art infrastructure models and generative AI systems, the following "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems" should be followed[76] .

Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems (excerpt)

I) Take appropriate measures to identify, assess, and mitigate risks throughout the development of advanced AI systems, including prior to their implementation and market launch, in order to identify, assess, and mitigate risks across the entire AI lifecycle.

- This includes employing a variety of internal and independent external testing measures in combination with assessment methods such as red teaming, and implementing appropriate mitigation measures to address identified risks and vulnerabilities. Testing and mitigation measures should, for example, aim to ensure the reliability, safety, and security of the system throughout its lifecycle so that the system does not pose unreasonable risks. To support such testing, developers should seek to enable traceability in relation to data sets, processes, and decisions made during system development. These measures should be documented and supported by technical documentation that is regularly updated.
- Such testing should be conducted in a secure environment to identify risks and vulnerabilities and to inform actions to address security, safety, social, and other risks, whether accidental or intentional, and should be conducted at several checkpoints throughout the AI lifecycle, particularly pre-deployment and pre-market It should be conducted at several checkpoints throughout the AI lifecycle, particularly at pre-deployment and pre-market. In designing and implementing testing measures, the organization commits to paying appropriate attention to the following risks
  - Chemical, biological, radiological, and nuclear risks, including how advanced AI systems can lower barriers to entry into weapons development, design acquisition, and use, including by non-state actors.
  - Offensive cyber capabilities are methods by which a system can find, exploit, or make operational use of vulnerabilities, etc., keeping in mind that there may be useful defensive applications of such capabilities that may be appropriate for inclusion in the system.
  - Risks to health and/or safety. Includes the impact of system interactions and the use of tools, including, for example, the ability to control physical systems or interfere with critical infrastructure.
  - Risks of models making copies of themselves, "self-replicating," or training other models.
  - Social risks and risks to individuals and communities, such as the potential for advanced AI systems and models to generate harmful bias and discrimination, or to violate applicable legal frameworks, including laws and regulations related to privacy and data protection.
  - Threats to democratic values and human rights, such as the promotion of disinformation and invasion of privacy.
  - The risk that a specific event will set off a chain reaction that will have a significant negative impact on the entire city, the entire territorial activity, and even the

[76] For the full text, see the G7 Leaders' Statement on the Hiroshima AI Process, "Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems" (October 2023), https://www.mofa.go.jp/mofaj/files/100573472.pdf Note that this is a living document that builds on existing OECD AI principles in response to recent developments in advanced AI systems. Advanced AI systems are also defined as the most advanced AI systems, including state-of-the-art infrastructure models and generative AI systems.

entire community.

- Each organization is committed to working with stakeholders across sectors to assess and adopt mitigation measures to address these risks, particularly systemic risks.
- Organizations committed to these commitments should also seek to promote research and investment in the security, safety, bias and disinformation, impartiality, accountability and interpretability, and transparency of advanced AI systems to increase their robustness and reliability against abuse.

II) Post-deployment, including market launch, identify and mitigate vulnerabilities and, if necessary, exploited incidents and patterns.

- Organizations should use AI systems as intended, at appropriate times commensurate with the level of risk, and monitor vulnerabilities, incidents, emerging risks, and exploits after implementation and take appropriate steps to address them. Organizations are encouraged to consider encouraging third parties and users to discover and report problems and vulnerabilities after implementation, for example, through incentive schemes, contests, or prizes to incentivize responsible disclosure of weaknesses. Organizations are further encouraged to work with other stakeholders to maintain proper documentation of reported incidents and mitigate identified risks and vulnerabilities. Where appropriate, mechanisms for reporting vulnerabilities should be available to a diverse set of stakeholders.

III) Contribute to improved accountability by publicizing the capabilities, limitations, and areas of appropriate and inappropriate use of advanced AI systems and helping to ensure adequate transparency.

- This should include publishing a transparency report with meaningful information on all significant new public announcements of advanced AI systems.
- These reports, instructions for use, and related technical documents should be kept up to date as appropriate and should include, for example
  - Details of assessments conducted on potential safety, security, social and human rights risks.
  - Critical limitations in model/system capability and performance that affect the appropriate use areas.
  - Discussion and evaluation of the safety and social impacts and risks of the model and system, including harmful bias, discrimination, threats to privacy violations, and impacts on equity.
  - Results of red teaming conducted to assess model/system fit after the development phase
- Organizations should ensure that the information within the transparency report is sufficiently clear and understandable to allow appropriate and relevant implementers and users to interpret the output of the model/system and allow users to use it appropriately. The transparency report should also be supported and provided by a robust documentation process, such as technical documentation and instructions for use.

IV) Work toward responsible information sharing and incident reporting among organizations developing advanced AI systems, including industry, government, civil society, and academia

- This includes, but is not limited to, responsible sharing of appropriate information, including but not limited to assessment reports, information on security and safety risks, dangerous intentional or unintentional capabilities, and attempts by AI stakeholders to circumvent safeguards throughout the AI life cycle.
- Each organization should establish or participate in mechanisms to develop, promote, and where appropriate, adopt shared standards, tools, mechanisms, and best practices to ensure the safety, security, and reliability of advanced AI systems.
- This should include ensuring adequate documentation and transparency throughout the entire AI lifecycle, especially with regard to advanced AI systems that pose significant

risks to safety and society.

- With a view to improving the safety, security, and reliability of advanced AI systems, organizations should collaborate with other organizations throughout the entire AI lifecycle to share relevant information and report to society. Organizations should also collaborate with relevant public authorities to share the aforementioned information, as appropriate.
- Such reporting should protect intellectual property rights.

V) Develop, implement, and disclose AI governance and risk management policies based on a risk-based approach, including privacy policies and mitigation measures

- Organizations should implement appropriate organizational mechanisms to develop, disclose, and implement risk management and governance policies, including, where feasible, accountability and governance processes to identify, assess, prevent, and address risks throughout the AI lifecycle.
- This includes disclosing privacy policies where appropriate, including for personal data, user prompts, and the output of advanced AI systems. Organizations are expected to follow a risk-based approach to establish and disclose AI governance policies and organizational mechanisms for implementing these policies. This should include accountability and governance processes to assess and mitigate risk where feasible throughout the AI lifecycle.
- Risk management policies should be developed according to a risk-based approach and a risk management framework should be applied throughout the lifecycle of the AI as appropriate and relevant to address the various risks associated with the AI system, and policies should be updated on a regular basis.
- The organization should establish policies, procedures, and training to ensure that personnel are familiar with their responsibilities and the organization's risk management practices.

VI) Invest in and implement robust security controls, including physical security, cyber security, and safeguards against insider threats throughout the AI lifecycle.

- This includes protecting model weights, algorithms, servers, and data sets through operational security measures for information security, appropriate cyber/physical access controls, etc.
- It also includes conducting cybersecurity risk assessments and implementing cybersecurity policies and appropriate technical and institutional solutions to ensure that the cybersecurity of advanced AI systems is adequate in light of the relevant environment and associated risks. Organizations should also take steps to reduce both the risk of unauthorized disclosure and the risk of unauthorized access by requiring that the storage and work on advanced AI system model weights be conducted in an appropriate and secure environment with restricted access. This includes a commitment to implement vulnerability management processes and regularly review security measures to ensure that they are maintained to a high standard and remain adequate to address risks.
- This would further include establishing a robust insider threat detection program consistent with protections for the most valuable intellectual property and trade secrets, for example, limiting access to non-public model weights.

VII) Where technically feasible, develop and implement reliable content authentication and comestion mechanisms, such as watermarking or other techniques, to enable users to identify AI-generated content.

- This includes, where appropriate and technically feasible, content authentication and provenance mechanisms for content created by the organization's advanced AI system. The provenance data should include the identifier of the service or model that created the content, but need not include user information. Organizations should also work to develop tools and APIs that allow users to determine, e.g., through watermarks, whether a particular piece of content was created by an advanced AI system.

Organizations should collaborate and invest in research, as appropriate, to advance the state of the art in this area.

- Organizations are further encouraged to implement other mechanisms, such as labeling or disclaimers, where possible and appropriate, so that users know they are interacting with an AI system.

VIII) Prioritize research to mitigate social, safety, and security risks and prioritize investments in effective mitigation measures.

- This includes investments in conducting, collaborating on, and investing in research and developing appropriate mitigation tools to help improve the safety, security, and reliability of AI and address key risks.
- Organizations to conduct, collaborate on, and invest in priority research that supports improving the safety, security, and reliability of AI and addresses critical risks, including maintaining democratic values, respecting human rights, protecting children and vulnerable populations, protecting intellectual property and privacy, and avoiding harmful bias, false and misinformation, and information manipulation Commitment. The organization will also commit to investing in the development of appropriate mitigation tools and will work to ensure that the risks of advanced AI systems, including environmental and climate impacts, are proactively managed and their benefits realized.
- Organizations are encouraged to share research and best practices in risk mitigation.

IX) Prioritize the development of advanced AI systems to address the world's greatest challenges, especially (but not limited to) the climate crisis, global health, education, etc.

- These efforts are designed to support progress on the UN Sustainable Development Goals and encourage the development of AI for global benefit.
- Organizations should prioritize responsible stewardship of reliable, human-centered AI and support digital literacy initiatives that enable individuals and communities to benefit from the use of advanced AI systems and to better understand the nature, capabilities, limitations, and impacts of these technologies Education and training of the general public, including students and workers, should be promoted to facilitate this process. Organizations should work with civil society and community groups to identify priority issues and develop innovative solutions to address the world's biggest challenges.

X) Promote the development of international technical standards and their adoption where appropriate.

- Organizations are also encouraged to contribute to and, where appropriate, utilize the development of international technical standards and best practices, including watermarking, and to work with standards development organizations (SDOs) when developing organizational testing methods, content authentication and provenance mechanisms, cybersecurity policies, public reporting, and other measures are encouraged to do so. In particular, they are encouraged to work on the development of interoperable international technical standards and frameworks that enable users to distinguish between AI-generated and non-AI-generated content.

XI) Protect personal data and intellectual property by implementing appropriate data input measures.

- Organizations are encouraged to take appropriate steps to control data quality, such as training data and data collection, to reduce harmful bias.
- Appropriate measures include transparency, privacy-preserving training techniques, and/or testing and fine-tuning to ensure that systems do not leak sensitive or confidential data.
- Organizations are encouraged to implement appropriate safeguards to respect privacy and intellectual property rights, including copyrighted content.
- Organizations should also comply with the applicable legal framework.

Appendix 4. for AI providers

In this chapter, first, "points" and "specific methods" of the contents described in "Part 4: Matters Concerning AI Providers" of this volume are explained. Then, specific methods that AI providers should be particularly aware of are explained among the "C. Common Guidelines" in "Part 2: Society to be Aimed by AI and Matters to be Tackled by Each Entity" of this volume.

Note that the "specific methods" described here are only examples. Some are written for both conventional and generative AI, while others apply to only one of them. In considering specific measures, it is important to take into account the degree and probability of risks posed by the AI system to be provided, technical characteristics, and resource constraints of each entity.

In addition, AI providers who deal with advanced AI systems should comply with Ⅰ) to ⑪) to ⅫⅢ) to the appropriate extent, referring to the description in "D. Guidelines common to business operators related to advanced AI systems" in "Part 2: Society to be aimed by AI and what each entity should work on" of this volume.

A. Explanation of "Part 4: Matters Related to AI Providers"

[Description in this volume (reprinted)]

- When AI system is implemented

(P-2) i. Risk measures that consider human life, body, property, spirit and environment

- ✧ Ensure that the AI system can maintain its performance level not only under the conditions of use expected at the time of provision, but also under a variety of conditions to avoid harm to the life, body, property, mind and environment of relevant stakeholders, including AI users, and to minimize risks (e.g., loss of control of interlocking robots). Consider methods (e.g., guardrail technology) to minimize risks (e.g., loss of control or inappropriate output) ("2) Safety").

[Point]

It is expected that AI will not cause harm to human lives, bodies, or property through actuators, etc., by taking countermeasures as necessary based on information from AI developers, etc., in consideration of the nature and manner of possible damage.

It is expected that measures to be taken when an AI causes harm to human life, body, or property through actuators, etc., should be organized in advance. In addition, necessary information on such measures is expected to be provided to AI users and users outside the business.

In addition to compliance with existing laws, regulations, and guidelines, it is also important to think of new technologies to deal with problems that new technologies cause.

[Specific Methods]

- Consideration for human disadvantage
  - ➢ Consideration for possible disadvantages to individuals (e.g., when profiling using AI in areas that may have a significant impact on the rights and interests of individuals. The following are examples of disadvantages that require consideration)
    - ✧ Incorrect decisions due to factual discrepancies in profiling results
    - ✧ Under- or overestimation of individuals due to the fact that only certain characteristics of an individual are used in profiling
    - ✧ If some of the profiling results for an individual share characteristics with a particular group, a negative decision for that group may result in a negative decision for that individual as well.
    - ✧ The profiling results in treatment that undermines a person's rights and interests, such as promoting unfair discrimination against a particular individual or group.

- ✧ Negative judgments enter into the process of predicting an uncertain future based on profiling results.
- ✧ The identification of an anonymous individual by comparing the results of profiling based on information about the anonymous individual with the results of profiling based on information about a specific individual.

- Incident Prevention
  - ➢ Establishment of a system that can ensure safety throughout the AI system (realization of fail-safe)
  - ➢ Promptly notify the AI developer of the existence of risks of which the AI developer is unaware, and discuss and consider countermeasures
  - ➢ Human involvement in safety checks, etc. prior to and during operation, and consideration of measures to prevent recurrence after the fact
  - ➢ Confirmation of the reliability of AI users through appropriate use declarations on the part of AI users, etc.

[Ref.]
- Tentative Discussion Paper on AI, Cabinet Office (May 2023)
- The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (October 2023)

[Description in this volume (reprinted)]
- When AI system is implemented

(P-2) ii. Provision that contributes to appropriate use

  - ✧ Properly define the points to be considered in the use of AI systems and services ("2) Safety")
  - ✧ Utilize AI within the limits set by the AI developer ("2) Safety")
  - ✧ Ensure data accuracy, up-to-dateness when necessary (that data is appropriate), etc. at the time of provision ("2) Safety")
  - ✧ Examine whether there are any differences between the assumed use environment of AI set by AI developers and the use environment of users of AI systems and services ("2) Safety").

- After providing AI systems and services
  - ✧ Regularly verify that AI systems and services are being used for appropriate purposes ("2) Safety")

[Point]

In providing AI systems and services, AI providers are expected to cooperate with relevant stakeholders to take preventive measures and post-response actions (e.g., information sharing, suspension/restoration, clarification of causes, and measures to prevent recurrence) according to the nature and manner of damage caused or resulting from incidents, security breaches, privacy violations, etc., that may or have occurred as a result of AI applications. (e.g., information sharing, suspension/restoration, clarification of causes, and measures to prevent recurrence) in cooperation with relevant stakeholders.

[Specific Methods]
- Cooperation with stakeholders and preventive and follow-up actions
  - ➢ Provision of information for the use of AI in an appropriate scope and manner
  - ➢ Prepare a list of response items, procedures, etc., for measures to be taken if an AI causes harm to human life, body, or property.
  - ➢ Implementation of measures to be taken in the event of a security breach
  - ➢ Implementation of measures to be taken in the event of an invasion of personal privacy
  - ➢ Share information with stakeholders when new risks are identified
  - ➢ Educational activities for society in general, including potential users
  - ➢ Periodic confirmation of appropriate use

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)

[Description in this volume (reprinted)]
- When AI system is implemented

(P-3) i. Consideration for biases in the composition and data of AI systems and services

- ✧ Ensure fairness of data at the point of provision, and consider bias of information to be referenced, external services to be linked, etc. ("3) Fairness")
- ✧ Regularly evaluate the inputs/outputs of AI models and the basis for decisions, and monitor the occurrence of biases. Also, if necessary, encourage AI developers to re-evaluate the bias of each technical element that constitutes the AI model and make decisions to improve the AI model based on the evaluation results ("3) Fairness").
- ✧ Consider the possibility that AI systems/services and user interfaces that receive the output results of AI models may include biases that arbitrarily limit business processes and the judgment of AI users and non-business users ("3) Fairness").

[Point]

AI providers are expected to be mindful of the possibility of bias in the judgments of AI systems and services, and to ensure that individuals and groups are not unfairly discriminated against based on the judgments of AI systems and services.
(Note: It is important to note that "impartiality" has multiple criteria, including group and individual impartiality.

[Specific Methods]

- Note that various biases determine the output of AI
  - ➢ Representativeness of data Bias by[77]
    - ✧ Potential for bias due to lack of data representativeness
    - ✧ Potential for bias by using data with inherent social bias
    - ✧ Pre-processing methods may cause unintended bias in input data at the time of use
  - ➢ Handling of personal information contained in data
    - ✧ When a large amount of data containing personal information is to be collected to meet data representativeness, it is handled with consideration for privacy, such as by masking or deleting personal information.
  - ➢ Algorithmic Bias
    - ✧ Depending on the algorithm, bias may occur due to sensitive attributes (personal attributes such as gender and race of the subject that should be excluded from the perspective of fairness)
  - ➢ Clarification of sensitive attributes
    - ✧ In articulating such attributes, consider the grounds enumerated in Article 14(1) of the Constitution and the attributes referred to in international human rights rules
  - ➢ Clarification of the content of fairness to be ensured with respect to sensitive attributes
  - ➢ Adding constraints to machine learning algorithms that meet fairness criteria
  - ➢ Use of tools (software) to check for bias
- Identification of criteria for impartiality (see "Column 10: Group and Individual Impartiality")
  - ➢ Criteria for group equity (example criteria below)
    - ✧ Remove sensitive attributes and make predictions based only on non-sensitive attributes (unawareness)
    - ✧ Ensure the same prediction results across multiple groups with different values of sensitive attributes (demographic parity)

[77] The data obtained by measurement or other means are not biased in part and are considered appropriate to represent the population as a whole.

- ✧ Adjust the ratio of the error of the predicted result to the actual result independent of the value of the sensitive attribute (equalized odds)
- ➢ Individual Equity Criteria (example criteria below)
  - ✧ Give the same prediction result for each individual with equal attribute values except for sensitive attributes
  - ✧ Give similar predicted results to individuals with similar attribute values (fairness through awareness)

Column 10: Group equity and individual equity

In general, the fairness requirement is concerned with the treatment of attributes that may cause "unfairness" such as race and gender (attributes requiring special consideration, which are synonymous with sensitive attributes). In this case, group fairness is to avoid discrimination (e.g., disadvantageous treatment of women) among different groups with respect to certain sensitive attribute values, while individual fairness is to avoid discrimination among "similar people" without necessarily being limited to classification by such specific attributes. At present, most of the machine learning elements that can be used universally for fairness evaluation (metrics) and measures are based on the assumption of group fairness, which requires "attributes that require consideration," and unless otherwise specified, they are based on the group fairness perspective. When "legitimacy" for the individual concerned is required, as mentioned above, the individual fairness perspective is required, so it is difficult to define general metrics, and measures to satisfy the requirement must be considered for each AI system. As for individual fairness, research on "degree of similarity" using distance learning has been proposed, and we look forward to future research on this topic.

- Achievement of equity through a qualitative approach/quantitative methods process (see "Column 11: Processes Related to Ensuring Equity Quality")
  - ➢ A risk analysis approach that qualitatively handles social demands, data on quality when using AI systems and services, etc., and considers the occurrence of lack of fairness as a risk.
    - ✧ First require qualitative fairness assurance, but if necessary, use quantitative fairness metrics such as "equality of results" from the implementation stage, and incorporate quantitative approaches as the content of the system and AI elements become more concrete.
    - ✧ It attempts to ensure the appropriateness of setting and selecting fairness metrics through analysis and design approaches, and is considered similar to the risk analysis-based approach in realizing "risk averseness," functional safety, etc. (see "Figure 19. Example of Process Structure for Ensuring Fairness Quality").

[Ref.]
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

Column 11: Process for ensuring fairness quality

Figure 19: Example of Process Structure for Ensuring Fairness Quality" shows an example of a series of processes to realize "equality of treatment" at a high level of abstraction, such as social requirements and quality at the time of use, qualitatively, and to embody the occurrence of unfairness (synonymous with loss of fairness) through a risk analysis approach that considers it as a risk, and to realize it from any stage of development as necessary through quantitative fairness metrics such as "equality of results. The diagram shows an example of a series of processes to achieve "equality of results" and other quantitative fairness metrics from any stage of development. This diagram is not intended to be a binding model for individual development concepts or stage settings, but rather a model to organize the flow from a qualitative approach to a quantitative method.

① The most abstract fairness requirement: at the level of "justice" or "human rights," it is required to be "equal" or "equal treatment.

② Social demand: "Equal treatment" is required at the level of social rules such as legal system or implicit ethical behavior, and "equal results" is required in the form of numerical targets, etc.

③ (iv) System requirements/requirements for AI elements: Either numerical equality of results or equality of treatment is required for goal setting as a level corresponding to the overall system design and the design of the machine learning elements.

⑤ Internal quality study: Even within the process of building a partial system of internal quality, either numerical equality of results or equality of treatment is required in the setting of goals.

⑥ Internal quality realization: At the level of internal quality, corresponding to the means of checking quality, there are possible ways to analyze the statistical distribution of results, to monitor statistical and analytical indicators other than the result distribution, and to explain the equality of treatment from the logical structure of the implementation.

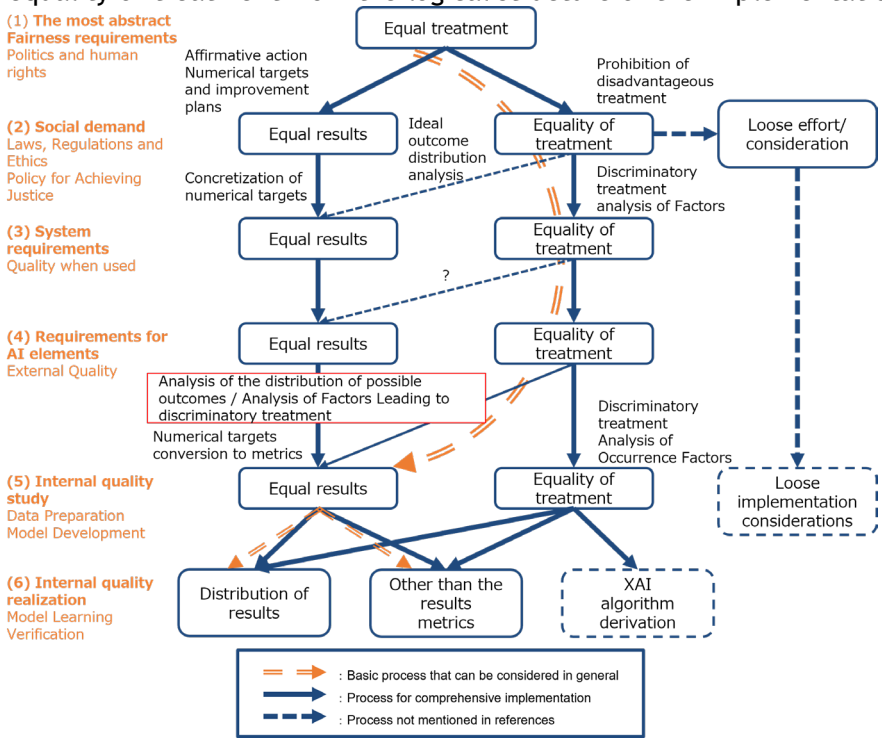

Figure 15 . Examples of Process Structures Related to Ensuring Fairness Quality

[Description in this volume (reprinted)]
- When AI system is implemented

(P-4) i. Introduction of mechanisms and measures to protect privacy

✧ Take measures to protect privacy through the process of implementing AI systems, including the introduction of mechanisms to control and restrict access to personal information appropriately in light of the characteristics of the technology to be employed (privacy by design) ("4) Privacy Protection").

[Point]

Privacy by Design should be broadly applied to all three aspects.

1) IT Systems
2) Responsible Business Practices
3) Physical design and network infrastructure

The goal of Privacy by Design, "to ensure privacy and to gain sustainable competitive advantages for the organization." can be achieved by implementing the following "Seven Basic Principles of Privacy by Design".

① Proactive, not ex post facto; preventive, not remedial
② Privacy as a default setting
③ Privacy built into the design
④ Holistic - positive-sum rather than zero-sum (an approach that yields all legitimate benefits & goals rather than a zero-sum approach that creates trade-offs)
⑤ Security from start to finish - all lifecycle protection
⑥ Visibility and Transparency - Maintaining Openness
⑦ Respect the privacy of users - Maintain a user-centered approach

[Specific Methods]

- Implementation of privacy measures based on Privacy by Design
  - ➢ Implementation of quality management (see "Table 6. Summary of Quality Management Implementation Items")
    - ✧ Ensure that data requiring protection is in compliance with applicable laws and regulations
    - ✧ Identify personal information requiring special consideration as defined by law
    - ✧ Determine reusable deliverables
    - ✧ Review consent arrangements with data providers and handle data in accordance with the arrangements

- Respect the privacy of relevant stakeholders and individuals
  - Deletion of information that violates personal privacy, updating of AI algorithms, etc. (when information that violates the privacy of AI users and other relevant stakeholders and individuals is obtained)
  - Requests for erasure of information that violates personal privacy, updates to AI algorithms, etc. (when information that violates the privacy of AI users and other relevant stakeholders and individuals is disseminated)

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- Ann Cavoukian, "Privacy by Design: The 7 Foundational Principles" (September 2011)

[Description in this volume (reprinted)]
- When AI system is implemented

(P-5) i. Introduction of mechanisms for security measures

✧ Throughout the process of providing AI systems and services, take appropriate security measures in light of the characteristics of the technologies employed (security by design) ("5) Ensuring Security").

[Point]

It is expected to pay attention to the security of AI and take reasonable measures to ensure the confidentiality, integrity, and availability of AI systems, in light of the technical level at that time. In addition, measures to be taken in the event of a security breach are expected to be organized in advance, taking into account the applications and characteristics of the AI system in question, the magnitude of the impact of the breach, and other factors.

The security of the AI system to be developed should be ensured by considering security from the early stage of the development process, referring to Security by Design, etc. defined by the Cabinet Cyber Security Center (NISC) in "Measures to incorporate information security from the planning and design stage. Ensure the security of the AI system to be developed by considering security from the early stage of the development process. Adding security functions after the fact or implementing security tools just before shipment may cause frequent rework, resulting in high development costs. Implementing security measures at an early stage of development will reduce rework and lead to the implementation and provision of an AI system with good maintainability.

[Specific Methods]

- Implement security measures based on Security by Design
  - Conduct Threat Assessment
    - ✧ Identify the threats and possible attacks that AI systems face; identify "what to protect" AI systems from.
  - Definition of Security Requirements
    - ✧ Defines the secure behavior of the AI system itself. Types of requirements include those related to system functions, usability, maintainability, performance, etc. Security requirements are the requirements related to security among the system requirements, and define the necessary targets for safe operation of the system. Describe security requirements as a part of the system requirement definition document or as a security requirement definition document.
  - Selecting a Security Architecture (Security Architecture)
    - ✧ Based on the architectural information required for AI systems provided by AI developers.
    - ✧ Customize and use the architecture recommended by the provider of the platform on which the AI system is installed, rather than developing your own architecture.
- Classification of attacks against AI
  - System malfunction
    - ✧ Examples of damage from reduced risk averseness include object detection failures in automated driving, missed driver anomalies in driver assistance, missed malware detection in information security measures, intrusion detection failures in security systems, abnormal behavior detection failures, and increased false positives and false negatives in pathology diagnosis systems. False positives and false negatives in pathological diagnosis systems are examples of damage.
    - ✧ Examples of damage from poor performance of AI systems include: decreased efficiency of dispatch allocation in transportation and logistics, increased traffic congestion and logistics costs; decreased correctness of product recommendations, demand forecasting, and store situational awareness in the retail sector; and decreased adequacy of admissions, hiring, and staffing.

        - ✧ Examples of damage caused by reduced fairness include unfair and discriminatory lending through credit screening systems; unfair and discriminatory admission, hiring, and staffing through human resource evaluation systems; and unfair and discriminatory criminal risk assessment through crime prevention systems.
    - ➢ Leakage of AI model information
        - ✧ An attack that leaks non-public information such as parameters and functions of an AI model may result in the leakage of trade secrets and other information related to the functionality of the AI model.
    - ➢ Leakage of sensitive information contained in training data
        - ✧ If sensitive information is included in training data, an attack that leaks training data information may result in invasion of privacy, disclosure of trade secrets, or violation of laws, regulations, or contracts.
        - ✧ If the training data set contains personal information such as medical information, customer sales information, or image data of military installations where photography is prohibited, an attack that leaks the information in the training data could cause damage to individuals (see "Table 4. Examples of Damage and Threats to Machine Learning Application Systems").
- Consideration of measures to be taken in the event of a security breach
    - ➢ initial response measures
        - ✧ Recovery through rollback of AI systems, use of alternative systems, etc.
        - ✧ Deactivation of AI system (kill switch)
        - ✧ Disconnection of AI systems from the network
        - ✧ Confirmation of the nature of the security breach
        - ✧ Reporting to relevant stakeholders
    - ➢ Use of insurance to facilitate compensation, indemnification, etc.
    - ➢ Establishment of a third-party organization to investigate, analyze, and make recommendations on the cause of the problem

[Ref.]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- Information-technology Promotion Agency, Japan "Security by Design Installation Guide" (August 2022)
- NCSC, "Guidelines for secure AI system development" (November 2023).

[Description in this volume (reprinted)]
- When AI system is implemented

(P-6) i. Documentation of system architecture, etc.

✧ To improve traceability and transparency, document the system architecture, data processing processes, etc. of the provided AI systems and services that influence decision-making ("6) Transparency").

[Point]

To ensure accountability of AI inputs, outputs, etc., AI providers should record and store logs of AI system inputs, outputs, etc., and document them with interpretable content to facilitate improvement of the process itself and to enhance communication and dialogue with relevant stakeholders. If necessary, make risk management documentation publicly available. Documentation increases transparency and allows for a human review process, thereby ensuring accountability.

[Specific Methods]
- Ensuring Accountability
  - Logging and storage of logs for AI systems and services
    - ✧ Purpose of logging and storage (e.g., whether the purpose is to determine the cause or prevent recurrence of incidents in areas that could endanger human life, limb, or property)
    - ✧ Frequency of logging, logging accuracy, and log storage period
    - ✧ Protect logs (ensure confidentiality, integrity, availability, etc.)
    - ✧ Capacity of log storage location
    - ✧ Log time (e.g., ensure accuracy by synchronizing time)
    - ✧ Scope of logs to be disclosed
    - ✧ Log storage method (e.g., in server, on storage media, etc.)
    - ✧ Log storage location (local or cloud, etc.)
    - ✧ Procedure for checking logs (how to access logs, etc.)
  - Adoption of AI systems that implement interpretable algorithms
    - ✧ The AI system to be used will employ a highly readable and interpretable AI model in advance
  - Adoption of technical methods to explain the results of algorithmic decisions to a certain degree
    - ✧ Global explanatory methods, such as "making the AI prediction and recognition process readable," in which the explanation is replaced by an interpretable AI model (e.g., the inference results of the AI model are explained by instances (examples) in the dataset or data processed from those instances). (e.g., SHAP, etc.)
    - ✧ Local explanatory methods for presenting the basis of prediction for specific input, such as "presentation of important features," "presentation of important training data," and its "expression in natural language" (e.g., explaining the inference logic and basis of judgment based on rules such as "if Weighting of important factors that significantly affect inference, etc.)
  - Data History Management
    - ✧ Manage (data provenance) when, where, and for what purpose data used for AI training, etc. is collected.
  - Analysis of AI model input/output trends
    - ✧ Analyze the AI's output trends based on multiple input/output combinations to the AI (e.g., observing changes in output when input patterns are varied slightly, etc.)
  - Update technical documentation as appropriate

[Ref.]
- AI Product Quality Assurance Consortium "AI Product Quality Assurance Guidelines" (June 2023)

- OECD, "Advancing accountability in AI" (February 2023).

[Description in this volume (reprinted)]
- After providing AI systems and services

(P-4) ii. Measures against invasion of privacy

✧ Gather information on privacy violations in AI systems and services as appropriate, take appropriate action in the event of such violations, and consider ways to prevent recurrence ("4) Privacy Protection").

[Point]

Since privacy infringement is a matter of personal perception and social acceptability shifts with the passage of context and time, it is expected to constantly collect relevant information (market trends, technologies, systems, etc.). It is also expected to establish relationships with experts in privacy infringement (academics, consultants, lawyers, consumer groups, etc.) and consult with them as necessary. In addition, it is also important to take initial actions in the event of a privacy violation in the actual business, as well as post-actions such as damage remedies, clarification of the cause, consideration of measures to prevent recurrence, and improvement measures.

[Specific Methods]
- Response by privacy protection organization
  - Aggregation of various information on new business and service offerings from various divisions within the company (with the goal of finding any omission of risk of privacy violations manifesting themselves to consumers and society).
  - Initial response led by the Privacy Officer and subsequent post-response measures such as damage relief, clarification of causes and measures to prevent recurrence (in the event of a privacy breach)
  - Building relationships with various departments within the company
    - ✧ In addition to receiving a wide range of privacy-related consultations from business divisions and others, it is expected that the company will always be in contact with divisions that handle AI systems and services on a regular basis, for example, by actively encouraging them to share their awareness of the issues. It is important to create a system and environment in which new business and technology development divisions can freely consult with each other without being burdened with their own concerns.
  - Establish a privacy protection organization structure (examples of structure patterns are listed below)
    - ✧ No privacy protection organization, but a person responsible for each department that handles AI systems and services
    - ✧ Establish a privacy protection organization (concurrently) to collaborate with the department handling AI systems and services
    - ✧ Establish a (dedicated) privacy protection organization to work with departments handling AI systems and services

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)

[Description in this volume (reprinted)]
- After providing AI systems and services

(P-5) ii. Vulnerability Response

✧ Since many attack methods against AI systems and services have been created, the latest risks and the trends of points to be taken care of in each process of provisioning to cope with them will be reviewed. In addition, consider eliminating vulnerabilities ("5) Ensuring Security").

[Point]

It is important for AI providers to provide AI users and off-business users with security measures for AI systems and services they provide themselves, as well as to share information on past incidents.

AI providers are expected to be mindful of the risk of security vulnerabilities in managing, improving, and adjusting AI models. They are also expected to inform AI users and non-business users of the existence of such risks in advance.

In Threat Analysis, the threats and possible attacks that AI systems and services face are organized, and "what to protect against" is clarified.

[Specific Methods]

- Attention to risks related to vulnerabilities to AI models (examples of risks below)
  - Risk of AI models artificially malfunctioning by inputting data with minute variations that cannot be discerned by humans to data that AI models can accurately determine as a result of insufficient learning, etc. (e.g., Adversarial example attack)
  - Risk of incorrect learning in supervised learning by mixing data with incorrect labeling, etc.
  - Risk that AI models can be easily replicated
  - Risk of being able to reverse engineer the data used to train from the AI model.
- Countermeasures against various machine learning-specific attacks (see "Table 5. Examples of machine learning-specific threats, attack interfaces, attack execution phases, attackers, and attack methods")
  - data poisoning attack
    - ✧ Confirmation of the authenticity of the data set and the reliability of the data set collection and processing process
    - ✧ Use of data poisoning detection techniques in data sets
    - ✧ Use of techniques to improve the robustness of the data set against data poisoning (e.g., increasing the number of data to reduce the impact of poisoning)
    - ✧ Training with learning methods that are robust against data poisoning (e.g., random smoothing, ensemble learning, etc.)
    - ✧ Eliminate/reduce poisoning from trained models
    - ✧ Conventional security measures against vulnerabilities in development software and development environments
  - Manipulation of validation/test data
    - ✧ Verify the reliability of the data set collection and processing process
    - ✧ Conventional security measures against vulnerabilities in development software and development environments
  - Model Poisoning Attack
    - ✧ Verify the reliability of the process of learning and delivering AI models
    - ✧ Use of model poisoning detection techniques
    - ✧ Eliminate or reduce poisoning of pre-trained or AI models
    - ✧ Use of learning mechanisms to remove or reduce poisoning
    - ✧ Conventional security measures against vulnerabilities in development software and development environments
    - ✧ Conventional security measures against vulnerabilities in the system environment and operating system during operation
  - evasive attack

- ✧ Methods for improving and evaluating the robustness of AI models against hostile data
- ✧ Restrictions on inputs to the AI model (restrictions on access rights, number and frequency of accesses)
- ✧ Use of Hostile Data Detection Technology
- ✧ Use of several different AI models and systems together
- ✧ Technical measures to prevent and mitigate model extraction attacks

➢ model extraction attack
- ✧ Use of model extraction attack detection techniques
- ✧ Processing of AI model output information, etc.
- ✧ ensemble learning
- ✧ Use of model extraction risk assessment techniques

➢ Information leak attack on training data
- ✧ Privacy Protection Study
- ✧ Privacy Protected Data Generation

[Ref.]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)
- Information-technology Promotion Agency, Japan "Security by Design Installation Guide" (August 2022)

[Description in this volume (reprinted)]

- After providing AI systems and services

(P-6) ii. Provide information to relevant stakeholders

- ✧ Provide timely and appropriate information on the AI systems and services provided in a plain and accessible form, for example, on the following matters ((6) Transparency)
  - The fact that AI is being used, appropriate/inappropriate use, etc. ("6) Transparency")
  - Information on safety, including technical characteristics of the AI system/service to be provided, foreseeable risks that may arise from the results of its use and their mitigation measures ("2) Safety")
  - Potential for output or program changes due to learning, etc. of AI systems/services ("1) Human-centric")
  - Information on the operational status of AI systems and services, causes of and responses to problems, incident examples, etc. ("2) Safety")
  - Information on updates to the AI system, if any, and the reasons for such updates ("2) Safety")
  - Collection policy, learning methods, and implementation system for data to be trained by the AI model ("3) Fairness," "4) Privacy protection," and "5) Security assurance")

[Point]

Taking into account the social context in which AI is used, such as when AI is used in fields that may have a significant impact on individual rights and interests, AI providers are expected to ensure the explainability of AI output results in order to obtain AI users' sense of conviction and security, and to present evidence for AI behavior for this purpose. It is expected that the AI will be able to explain the results of its outputs. In doing so, it is expected to analyze and understand what kind of explanation is required, and take necessary measures.

Once risks have been assessed and addressed, it will be important to verify and share with relevant stakeholders whether the AI system complies with regulatory, AI governance, and ethical standards. This will facilitate an understanding of the risks and the rationale behind decisions and actions. In addition to establishing monitoring and review processes and tools, it is important to communicate and review regularly to ensure that information about undesirable AI behaviors and incidents is also shared with relevant stakeholders.

[Specific Methods]

- Sharing information about AI systems and services
  - The AI system/service to be provided is AI-based and its application/method
  - Benefits and risks according to the nature of AI and the manner of its use, etc.
  - Methods of periodic confirmation regarding the scope and method of utilization of AI systems and services to be provided (in particular, observation and confirmation methods when AI systems are updated autonomously), importance and frequency of confirmation, risks due to unconfirmed, etc.
  - AI system updates and AI system inspections, repairs, etc. conducted to improve AI functionality and control risks through the process of utilization.
  - Details of assessments conducted on risks to safety, security, social risks, and human rights
  - Appropriate use areas and the capacity and performance limitations of AI models and AI systems and services that affect their use
  - Discussion and evaluation of the safety and social implications and risks of AI models and AI systems and services, including harmful bias, discrimination, threats of privacy violations, and impacts on equity
  - Results of red teaming conducted to assess the suitability of AI models and AI systems and services after the development phase
  - Points to keep in mind when providing information

- ✧ AI users share necessary information at the right time
- ✧ Provide information to be provided about AI systems and services prior to use.
- ✧ If the above information cannot be provided prior to the use of AI systems and services, a system shall be established to respond to feedback from AI users and non-business users, in accordance with the risks assumed based on the nature of AI and the manner of its use.

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)
- NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023).
- OECD, "Advancing accountability in AI" (February 2023).

[Description in this volume (reprinted)]
- After providing AI systems and services

(P-7) i. Explanation of the status of compliance with "common guidelines" for AI users

- ✧ Encourage AI users to use AI appropriately and provide the following information to AI users ("7) Accountability")
  - Reminder about the use of data whose accuracy and, where necessary, up-to-dateness (that the data is appropriate), etc. are ensured ("2) Safety")
  - Alert against inappropriate AI model learning through in-context learning ("2) Safety")
  - Points to keep in mind when entering personal information ("4) Privacy Protection")
- ✧ Alerting users about inappropriate input of personal information into AI systems and services provided ("4) Privacy Protection").

[Point]

AI providers are expected to provide AI users and non-professional users with information and explanations about the characteristics of AI systems in light of the nature and purpose of the AI they use, according to the amount of knowledge and ability they have, and to engage in dialogues with various stakeholders, in order to gain trust in AI from people and society, taking into account the purposes of other "common guidelines". AI users and non-business users are expected to fulfill reasonable accountability by providing information and explaining the characteristics of AI systems to AI users and non-business users in light of the nature and purpose of the AI they use, and by engaging in dialogue with various stakeholders.

When AI is used in a field that may cause harm to human life, body, or property, AI providers are expected to take measures as necessary based on information from AI developers, etc., taking into account the nature and manner of possible harm, and to explain the details of such measures to AI users and users outside the business to a reasonable extent.

[Specific Methods]
- Alerting AI users and off-business users to the use of AI systems and services and related actions
  - ➢ Inspection and repair of AI and updating of AI systems, as well as promotion of responses to AI users and off-business users (the purpose is to ensure that AI does not cause harm to human life, body, or property through actuators, etc.). Provide timely and appropriate information and reminders of problems from the time problems are discovered until updates are provided).
  - ➢ Provision of information on measures to be taken in the event of harm to human life, body, or property (if necessary)
  - ➢ Confirmation by recording and storing logs of input/output, etc., and alerting users to inappropriate input (for the purpose of deterring malicious use by AI users and non-business users)

- Measures to be taken when the privacy of relevant stakeholders and individuals, such as AI users and non-business users, is violated.
- Appropriate scope and method of use of AI systems and services (information provided based on information and explanations provided by AI developers, etc., confirming the purpose, use, nature and capabilities of the AI).

[Ref.]

- The White House, "Blueprint for an AI Bill of Rights (Making Automated Systems Work for The American People)" (October 2022).

[Description in this volume (reprinted)]
- After providing AI systems and services

(P-7) ii. Documentation of terms of service, etc.

- ✧ Create terms of service for AI users and non-business users ("7) Accountability").
- ✧ Clearly state the privacy policy ("7) Accountability")

[Point]

In order to eliminate uncertainty in AI service provision and to maintain and manage appropriate service level, it is effective to utilize Service Level Agreement (SLA), which is a common recognition of assurance standards regarding service contents, scope, quality, etc. SLA is expected to clarify the scope, contents, and preconditions of services, as well as the required standards for service level, and to form a common recognition between AI users, non-professional users, and AI providers, The SLA is expected to clarify the scope, contents, and preconditions of services, as well as the required level of service level, and to form a common understanding among both AI users, non-business users, and AI providers.

In order to build a relationship of trust with AI users and non-business users and to secure society's confidence in business activities, it is expected to formulate and publish a "policy and approach to personal information protection (so-called privacy policy, privacy statement, etc.)".

[Specific Methods]
- Creation of Terms of Service
  - ➢ Targeted AI services and requirement levels
    - ✧ In setting up the system, determine the order of priority in consideration of the impact on business operations in the event of an incident, and define the system with a focus on the most important items.
    - ✧ When determining service levels, objective items (e.g., quantitative numerical values, measurement by mathematical formulas, etc.) shall be defined to prevent differences in perception between AI users or off-business users and AI providers.
- Formulation and publication of privacy policy and policy regarding the protection of personal information (privacy policy, privacy statement, etc.)
  - ➢ Clarification of outsourced processing
    - ✧ Promote transparency in outsourced processing, such as clarifying whether or not outsourcing is performed and the details of the outsourced operations.
  - ➢ official announcement
    - ✧ After the policy is formulated, publicize it by posting it on a website, etc., and explain it to the public in advance in an easy-to-understand manner.
- Update documentation as appropriate

[Ref.]
- Ministry of Economy, Trade and Industry "SLA Guidelines for SaaS" (January 2008)
- The White House, "Blueprint for an AI Bill of Rights (Making Automated Systems Work for The American People)" (October 2022).

B. Explanation of "Common Guiding Principles" in "Part 2"

This section describes specific methods that are not mentioned in "Part 4: Matters Related to AI Providers" of this volume, but are of particular importance to AI providers in the "Common Guidelines" in "Part 2" of this volume.

[Contents of this volume (reprinted)] * Only the columns are excerpted.

1) Human-centered

Each entity should ensure that the development, provision, and use of AI systems and services do not violate at least constitutionally guaranteed or internationally recognized human rights as a basis from which all matters to be addressed, including each of the items described below, are derived. It is also important to act in such a way that AI expands people's capabilities and enables the pursuit of diverse happiness (well-being) of diverse people.

- Related to 1) Human Dignity and Individual Autonomy
  [related description].
  - Respect human dignity and individual autonomy, taking into account the social context in which AI will be used

  [Specific Methods]
  - Promote research on social, safety, and security risk reduction
    - Invest in research and effective mitigation measures to reduce social, safety, and security risks (examples of research include)
      - Maintaining democratic values
      - Respect for Human Rights
      - Protection of children and vulnerable groups
      - Protection of Intellectual Property Rights and Privacy
      - Avoidance of harmful prejudice
      - Avoidance of false and misinformation
      - Avoidance of information manipulation, etc.
    - Research and share best practices on risk mitigation (to the extent possible)

- Related to "(2) Consideration of decision-making and emotional manipulation by AI
  [related description].
  - We will not develop, provide, or use AI systems or services for the purpose of, or on the premise of, improperly manipulating human decision-making, cognition, or other emotions.

  [Specific Methods]
  - Measures against decision-making, emotional manipulation, etc.
    - Alerting AI users and off-business users
    - Promote sharing and awareness of the existence of risks such as dependence on AI and manipulation of decision-making and emotions in educational settings, etc.
    - Consider incentives (e.g., reward programs, contests, products, etc.) to encourage post-introduction vulnerability discovery and reporting

- Related to "(3) Countermeasures against false information, etc.
  [related description].
  - We recognize that the risk of AI-generated disinformation, misinformation, and biased information destabilizing and confusing society is increasing, and we must take necessary countermeasures.

  [Specific Methods]

- Consideration of risk avoidance measures for false information, misinformation, biased information, etc.
  - Development and implementation of technologies, such as digital watermarking and other technologies, that enable AI users and off-business users to identify information as AI-generated.
  - Provide information literacy education for a wide range of ages

- Related to (4) Ensuring Diversity and Inclusion
  [related description].
  - In addition to ensuring fairness, care will be taken to facilitate the use of AI by the socially vulnerable so that more people can enjoy the benefits of AI without creating so-called "information and technology weaklings".

  [Specific Methods]
  - A commitment to AI utilization that leaves no one behind.
    - Improvement of UI (user interface) and UX (user experience)
    - Establishment of a safe and secure user environment
    - Development of public digital platforms

- Related to (6) Ensuring sustainability
  [related description].
  - In the development, provision, and use of AI systems and services, the impact on the global environment is also considered throughout their lifecycle.

  [Specific Methods]
  - Consideration of common global issues
    - Supporting progress on the UN Sustainable Development Goals and encouraging the development and use of AI for global benefit (examples of global issues below)
      - Climate change measures
      - Human Health and Well-Being (World Health)
      - quality education
      - Eradication of poverty, a world without hunger
      - Maintain sanitation
      - Affordable clean energy
      - Eradication of inequality
      - Responsible consumption and production, etc.

[Ref.]
- Ministry of Internal Affairs and Communications "White Paper on Information and Communications 2021" (July 2021)
- United Nations "Sustainable Development Goals" (September 2015)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(2) Safety

Each entity should ensure that the development, provision, and use of AI systems and services do not cause harm to the lives, bodies, or property of stakeholders. In addition, it is important to ensure that no harm is caused to the psyche and the environment.

- Consideration for human life, body, property, spirit and environment
  [related description].
  - Examine measures to be taken in the event of a situation that compromises the safety of AI systems and services, and prepare for prompt implementation in the event of such a situation.

  [Specific Methods]
  - Organize incident countermeasures and consider measures to be taken in the event of an incident
    - Organizing Incident Response
      - Preparation of a communication system in the event of a hazardous situation
      - Organize methods for investigating causes and restoration work
      - Examine measures to prevent recurrence and organize response policies
      - Set up a method for sharing information about the incident
    - initial response measures
      - Recovery through rollback of AI systems, use of alternative systems, etc.
      - Deactivation of AI system (kill switch)
      - Disconnection of AI systems from the network
      - Confirmation of the nature of the harm
      - Reporting to relevant stakeholders
    - Use of insurance to facilitate compensation, indemnification, etc.
    - Establishment of a third-party organization to investigate, analyze, and make recommendations on the cause of the problem

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(3) Fairness

In the development, provision, and use of AI systems and services, it is important for each entity to strive to eliminate unfair and harmful prejudice and discrimination against specific individuals or groups on the basis of race, gender, nationality, age, political beliefs, religion, and other diverse backgrounds. It is also important for each entity to develop, provide, and use AI systems and services after assessing whether such unavoidable biases are acceptable from the perspective of respecting human rights and diverse cultures, while recognizing that some biases still cannot be avoided.

- Related to "(2) Intervention of human judgment"
  [related description].
  - To ensure that AI output results are not unbiased, consider using AI not only to make decisions on its own, but also to intervene with human judgment.

  [Specific Methods]
  - Judgment regarding the need for human judgment intervention (examples of judgment criteria below)
    - The nature of the rights and interests of AI users and off-business users affected by the output of AI, and the intentions of AI users and off-business users.
    - Degree of reliability of the AI's output (superiority or inferiority to the reliability of human judgment)
    - Timeframe required for human decision making
    - Competencies expected of AI users and out-of-business users who make decisions
    - The necessity of protection of the subject of the decision (e.g., whether to respond to individual applications by humans or to mass applications by AI systems/services, etc.)
    - Uncertainty of statistical future predictions
    - Necessity and degree of convincing reasons for decisions (judgments)
    - Assumed degree of discrimination based on race, creed, and gender due to inclusion of social bias against minorities, etc. in the study data
  - Ensuring the effectiveness of human judgment
    - Clarify in advance the items on which humans should make judgments based on explanations obtained from the AI with explainability (when it is appropriate for humans to make final judgments on the AI's output).
    - Provide information and explanations so that AI users and off-business users can acquire the necessary abilities and knowledge to appropriately judge AI outputs (when it is appropriate for humans to make final judgments on AI outputs).
    - Organize responses in advance to ensure the effectiveness of human decisions.

[Ref.]
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(6) Transparency

In developing, providing, and using AI systems and services, it is important for each entity to provide information to stakeholders to the extent necessary and technically feasible and reasonable, while ensuring the verifiability of AI systems and services, taking into account the social context in which AI systems and services are used. The following is a brief overview of the key points of this process.

- Related to "(iv) Improvement of accountability and interpretability to relevant stakeholders
  [related description].
  - Analyze and understand what kind of explanations are required and take necessary actions in order to obtain the relevant stakeholders' sense of conviction and reassurance, and to present evidence for the AI's operation for this purpose.

  [Specific Methods]
  - Ensuring Accountability
    - Clarification of areas where explanations are lacking, taking into account the needs and opinions of AI users and non-business users, and examination of explanatory content in cooperation with AI developers.
    - Context analysis in collaboration with stakeholders, including AI developers, and research and documentation of potential risks (e.g., targets to be affected, situations in which impacts may occur, etc.)
    - Monitoring and review interfaces to track risk and ensure frequency, functionality and effectiveness of implementation
    - Establish and communicate redress mechanisms, including processes for stakeholders to raise grievances

[Ref.]
- NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023)
- OECD, "Advancing accountability in AI" (February 2023).

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(8) Education and literacy

Each entity is expected to provide the necessary education to ensure that those involved in AI within the entity have the knowledge, literacy, and sense of ethics to have a correct understanding of AI and socially correct use of AI. Each entity is also expected to educate its stakeholders, taking into account the characteristics of AI such as complexity and misinformation, as well as the possibility of intentional misuse.

[related description].
- Take necessary steps to ensure that those involved in AI within each entity have a sufficient level of AI literacy in their involvement
- As the segregation of AI and human work is expected to change with the expanded use of generative AI, education, reskilling, etc. will be considered to enable new ways of working.

[Specific Methods]
- Ensuring AI literacy
  - Develop an AI policy that clearly defines roles and responsibilities and disseminate it to those involved in AI within the entity.
  - Define and inform those involved in AI within the entity about the characteristics of a trustworthy AI
  - Collect information on laws and regulations applicable to AI systems and disseminate this information to those involved in AI within the entity.

  - Gathering and disseminating potential adverse effects that may arise from AI systems to those involved in AI within the entity
  - Informing those involved in AI within the entity that data and digital technologies, including AI, are used in a variety of operations
- Education and reskilling
  - Provide training that comprehensively addresses the technical and socio-technical aspects of AI risk management
  - Education to improve resilience to environmental changes, etc.
    - ✧ Flexible mental rotation between "vertical thinking" and "horizontal thinking
    - ✧ Strengthen modeling skills required for organizational assessment
    - ✧ Extending the learning horizon in skill areas that are weak in agile thinking
    - ✧ Improved valuation skills to analyze uncertainties that are difficult to predict with past experience and expertise
    - ✧ Conversion of "key points" of organizational evaluation (to "instantaneous + sustainable" mobility)
    - ✧ How to evaluate management resynchronization (Configuration, Architecture, Synthesis, Dissemination), which is becoming increasingly complex and sophisticated as forms of AI governance become increasingly hybrid, centralized and decentralized

[Ref.]
- Ministry of Economy, Trade and Industry, Information-technology Promotion Agency, Japan "Digital Skill Standards ver.1.1" (August 2023)
- Cabinet Office, "Principles for a Human-Centered AI Society" (March 2019)
- NIST, "AI Risk Management Framework Playbook" (January 2023).

[Contents of this volume (reprinted)] * Only the columns are excerpted.
10) Innovation
Each entity is expected to strive to contribute to the promotion of innovation throughout society.

[related description].
- Ensure interconnectivity and interoperability between own AI systems and services and other AI systems and services
- Conform to standard specifications, if any

[Specific Methods]
- Standardization of data formats, protocols, etc.
  - Data format (syntax and semantics) for AI input/output, etc.
  - Connection methods for coordination between AI systems and services (especially protocols at each layer if via networks)
  - Ensure that languages are consistent when implementing multiple AI models or utilizing new datasets. If different, consider adjustments such as tokenization methods, vocabulary expansion, etc.

[Ref.]
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

Appendix 5. for AI business users

In this chapter, first, "points" and "specific methods" of the contents described in "Part 5: Matters Concerning AI Users" of this volume are explained. After that, specific methods that AI users should be particularly aware of are explained among the "C. Common Guidelines" in "Part 2: Society to be Aimed by AI and Matters to be Tackled by Each Entity" of this volume.

Note that the "specific methods" described here are only examples. Some are written for both conventional and generative AI, while others apply to only one of them. In considering specific measures, it is important to take into account the degree and probability of risks posed by AI systems and services to be used, technological characteristics, and resource constraints of each entity.

In addition, AI users who deal with advanced AI systems should comply with Ⅰ) to ⑪) to Ⅻ) to the extent appropriate, referring to the description in "D. Guidelines common to businesses involved in advanced AI systems" in "Part 2: Society to be aimed by AI and what each entity should work on" of this Part.

A. Explanation of "Part 5: Matters Related to AI Business Users"

[Description in this volume (reprinted)]

- When using AI systems and services

(U-2) i. Appropriate use with safety in mind

- ✧ Use AI systems and services within the scope assumed by the AI provider in its design, in compliance with the usage considerations specified by the AI provider ("2) Safety").
- ✧ Input data that is accurate, up-to-date when necessary (data is appropriate), etc. ("2) Safety")
- ✧ Understand the degree of accuracy and risk with respect to AI output and check various risk factors before use ("2) Safety")

[Point]

AI users should use AI based on information provided by AI providers (including information from AI developers) and explanations, and should also take into account the social context in which AI is used.

In the utilization of AI operated through actuators, etc., if AI is scheduled to be shifted to human operation when certain conditions are met, AI users and off-business users are expected to be aware in advance of their responsibilities before, during, and after the transition. They are also expected to receive explanations from AI providers on transition conditions, transition methods, etc., and acquire the necessary capabilities and knowledge.

In using AI, it is important for AI users, based on information provided by AI providers (including information from AI developers), to cooperate with relevant stakeholders to take preventive and follow-up measures (e.g., information sharing, suspension/restoration, clarification of causes, and measures to prevent recurrence) according to the nature and manner of damage caused by incidents that may or have occurred through the use of AI, security breaches, privacy violations, and so on. It is important to cooperate with relevant stakeholders to take preventive and post-response measures (information sharing, suspension/restoration, clarification of causes, measures to prevent recurrence, etc.).

[Specific Methods]

- Obtaining information about AI systems and services
  - ➢ Appropriate uses and methods of AI systems and services to be used
  - ➢ Benefits and risks according to the nature of AI and the manner of its use, etc.

  - Methods of periodic confirmation regarding the scope and method of AI utilization (especially, observation and confirmation methods when AI is updated autonomously), importance and frequency of confirmation, risks due to unconfirmed, etc.
  - AI system updates and AI inspections and repairs, etc. conducted to improve AI functionality and control risks through the process of utilization.
- Use in an appropriate scope and manner
  - Recognition of benefits and risks according to the nature of AI, mode of use, etc., and understanding of appropriate uses (before use)
  - Acquisition of necessary knowledge and skills for proper use (before use)
  - Periodic checks on whether AI is being utilized in an appropriate scope and manner (during use)
  - Updating the AI system and inspecting and repairing the AI, or requesting the AI provider to perform these tasks (with the aim of improving the functionality of the AI and limiting risks through the process of utilization) (in use)
    - ✧ However, consider that updates may affect other AIs that work together.
  - Feedback of incident information to the AI provider (or AI developer through the AI provider) (when any incident occurs, including when there are signs that an incident may occur)
- Preventive and follow-up actions taken in cooperation with relevant stakeholders
  - Provision of information for use in an appropriate scope and manner
  - Implementation of measures to be taken when an AI causes harm to human life, body, or property
  - Implementation of measures to be taken in the event of a security breach
  - Implementation of measures to be taken in the event of an invasion of personal privacy
  - Educational activities for society in general, including potential users
  - Prompt sharing of information on incidents, etc., with AI providers and AI developers and consideration of countermeasures

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)

[Description in this volume (reprinted)]
- When using AI systems and services

(U-3) i. Consideration for bias in input data and prompts

✧ Input data in a manner that ensures impartiality to avoid significant lack of fairness, and make business use decisions on AI output results responsibly, paying attention to bias in the prompts ("3) Impartiality").

[Point]

AI users are expected to contact the AI provider (or the AI developer through the AI provider) as necessary if they have any doubts about the AI's output results.

Given that the output of AI may be determined by the data at the time of learning, AI users are expected to pay attention to the representativeness of data used for AI learning, etc. and social biases inherent in the data, depending on the social context in which AI is used.

In order to maintain fairness in the results of judgments made by the AI, AI users are expected to intervene with human judgment, such as whether or not to use the judgment, or how to use it, based on the social context in which the AI is used and people's reasonable expectations.

[Specific Methods]

- Note that the output of AI is determined by various biases (points to consider when deciding whether or not to contact the AI provider)
  - Bias due to representativeness of data
    - ✧ Potential for bias due to lack of data representativeness
    - ✧ Potential for bias by using data with inherent social bias
    - ✧ Pre-processing methods may cause unintended bias in input data at the time of use
  - Handling of personal information contained in data
    - ✧ When a large amount of data containing personal information is to be collected to meet data representativeness, it is handled with consideration for privacy, such as by masking or deleting personal information.
  - Algorithmic Bias
    - ✧ Depending on the algorithm, bias may occur due to sensitive attributes (personal attributes such as gender and race of the subject that should be excluded from the perspective of fairness)
  - Clarification of sensitive attributes
  - Clarification of the content of fairness to be ensured with respect to sensitive attributes
  - Adding constraints to machine learning algorithms that meet fairness criteria
- Identification of fairness criteria (see Column 10: Collective and Individual Fairness)
  - Identification of criteria for group equity (examples of criteria are provided below)
    - ✧ Remove sensitive attributes and make predictions based only on non-sensitive attributes (unawareness)
    - ✧ Ensure the same prediction results across multiple groups with different values of sensitive attributes (demographic parity)
    - ✧ Adjust the ratio of the error of the predicted result to the actual result independent of the value of the sensitive attribute (equalized odds)
  - Identification of example criteria for individual fairness (example criteria are listed below)
    - ✧ Give the same prediction result for each individual with equal attribute values except for sensitive attributes
    - ✧ Give similar predicted results to individuals with similar attribute values (fairness through awareness)

[Ref.]

- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

[Description in this volume (reprinted)]
- When using AI systems and services

(U-4) i. Measures against improper input of personal information and invasion of privacy

- ✧ Take precautions to ensure that personal information is not inappropriately entered into AI systems and services ("4) Privacy Protection").
- ✧ Gather information on privacy violations in AI systems and services as appropriate, and consider prevention ("4) Privacy Protection").

[Point]

With regard to the handling of personal information when using AI systems and services, personal information should be handled appropriately in accordance with the rules of the Personal Information Protection Law, referring also to the "Alert on the Use of Generated AI Services, etc." by the Personal Information Protection Commission.

By establishing a privacy protection organization, the company can foster close communication among departments that use AI systems and services, including new business departments within the company, gather relevant information from outside experts and others, and consider measures from multiple perspectives in a substantive manner. The scope of privacy protection considerations is expanding every day due to technological innovation and increasing consumer awareness of privacy. Therefore, it is important to establish a privacy protection organization that can ensure multifaceted consideration and prompt response to social needs, such as technological innovation and consumer awareness, with respect to privacy issues. In addition, when personal information of consumers is handled globally, sufficient consideration should be given to the application of laws and regulations of other countries and a global system should be established to address privacy protection.

[Specific Methods]

- Response by privacy protection organization
  - ➢ Aggregation of various information about new business and service offerings from various organizations within the company (with the goal of finding omissions of risks of privacy violations manifesting themselves to consumers and society).
  - ➢ Initial response led by the privacy officer and subsequent post-response measures such as damage relief, clarification of causes and measures to prevent recurrence (in the event of a privacy violation)
  - ➢ Building relationships with various organizations within the company
    - ✧ In addition to receiving a wide range of privacy-related consultations from various organizations, it is also expected that the AI system/service provider will always be in contact with organizations that use AI systems and services on a regular basis, for example, by actively encouraging them to share their awareness of the problems they face. It is important to create a system and environment where organizations that develop and use new businesses and technologies can freely consult with each other without being burdened with their problems.
  - ➢ Establish a privacy protection organization structure (examples of structure patterns are listed below)
    - ✧ No privacy protection organization, but a responsible person for each organization using AI systems and services
    - ✧ Establish a privacy protection organization (with concurrent duties) to work with organizations that use AI systems and services
    - ✧ Establish a (dedicated) privacy protection organization to work with organizations that use AI systems and services
- Preliminary organization and implementation of measures to be taken in the event of a privacy violation
  - ➢ Preliminary organization of measures to be taken in the event of a privacy violation

        - ✧ When information is provided by AI providers (including from AI developers) regarding measures to be taken in the case of violation of personal privacy, we will consider the measures to be taken with due attention.
    - ➢ Erase information that could lead to invasion of personal privacy, update AI algorithms, etc. (When information that could lead to invasion of personal privacy is obtained)
    - ➢ Requests to AI providers, etc. to delete information that may lead to invasion of personal privacy, requests to AI developers, AI providers, etc. to update AI algorithms, etc. (when information that may lead to invasion of personal privacy is obtained)
- Enter prompts containing personal information
    - ➢ For example, if the personal data to be entered in the use of the generated AI service is planned to be used by the provider of the generated AI service as training data for AI, care should be taken not to enter prompts that include personal data for which consent has not been obtained.
    - ➢ Attention to information to be input into AI
        - ✧ Avoid giving confidential information (including not only one's own information but also that of others) to the AI unnecessarily, for example, by overly empathizing with the AI.
    - ➢ Respect for privacy
        - ✧ Respect the privacy of individuals when collecting data to train AI for their own use.

[Ref.]
- Personal Information Protection Commission, "Alert on the Use of Generated AI Services, etc." (June 2023)
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)

[Description in this volume (reprinted)]
- When using AI systems and services

(U-5) i. Implementation of security measures

✧ Comply with security considerations by AI providers ("5) Ensuring Security")

[Point]

AI users are expected to keep in mind when using AI systems/services any information provided by AI providers (including information from AI developers) on measures to be taken in case of security breaches. In addition, AI users are expected to report to the AI provider (or to the AI developer through the AI provider) any security questions they may have in using AI systems/services.

AI users are expected to pay attention to the security of AI systems and take necessary security measures in cooperation with off-business users, based on information provided by AI providers (including information from AI developers), if security measures are expected to be implemented on the off-business users' side.

[Specific Methods]
- Perceived Risks Related to Vulnerabilities
  - Risk of AI models artificially malfunctioning by inputting data with minute variations that cannot be discerned by humans to data that AI models can accurately determine as a result of insufficient learning, etc. (e.g., Adversarial example attack)
  - Risk of incorrect learning by mixing incorrectly labeled data in supervised learning
  - Risk of AI models being easily replicated
  - Risk of reverse engineering of data used for training from AI models
- Consideration of measures to be taken in the event of a security breach
  - initial response measures
    - ✧ Recovery through rollback of AI systems, use of alternative systems, etc.
    - ✧ Deactivation of AI system (kill switch)
    - ✧ Disconnection of AI systems from the network
    - ✧ Confirmation of the nature of the security breach
    - ✧ Reporting to relevant stakeholders
  - Use of insurance to facilitate compensation, indemnification, etc.
  - Establishment of a third-party organization to investigate, analyze, and make recommendations on the cause of the problem

[Ref.]
- Information-technology Promotion Agency, Japan "AI Handbook for Security Professionals" (June 2022)
- Information-technology Promotion Agency, Japan "Security by Design Installation Guide" (August 2022)

[Description in this volume (reprinted)]
- When using AI systems and services

(U-6) i. Provide information to relevant stakeholders

✧ Input data in a manner that ensures impartiality to avoid significant lack of fairness, and obtain output results from the AI system/service with attention to bias in the prompts. Then, when the output results are used to make business decisions, the results are communicated to relevant stakeholders ("3) Fairness" and "6) Transparency").

[Point]

When AI users use AI in areas that may have a significant impact on the rights and interests of individuals, etc., AI users are expected to take into account the social context in which AI is used, and to obtain the conviction and peace of mind of users outside their work, and to provide evidence for the operation of AI for this purpose (i.e., AI systems are expected to predict, recommend, or provide evidence of the underlying decision-making factors and plain and understandable information about the decision-making process), etc., it is expected to ensure the explainability of AI output results. In doing so, it is expected to improve the explainability of AI output results by analyzing and understanding what kind of explanations are required of AI users in order to build and maintain individual trust, and by taking necessary measures.

[Specific Methods]
- Provide information to relevant stakeholders
  - Clarification of Explanatory Objectives
    - ✧ Clarification of the scope of nondisclosure through the conclusion of an agreement with the AI provider in the form of a limited scope of nondisclosure (including that set by the AI developer).
  - Methods of explaining AI systems and services prior to implementation and testing of the explanations themselves
  - Obtaining feedback on description
    - ✧ Obtain feedback from stakeholders, including non-business users, and potentially affected individuals and groups on the accuracy, clarity, etc. of explanations
  - Provide information on AI models
    - ✧ Include information on input data types and sources, high-level data transformation process, decision criteria and rationale, risks and mitigation measures, etc.
  - Points to keep in mind when providing information
    - ✧ Share necessary information with non-work users at the right time
    - ✧ Provide information to be provided about AI systems and services prior to use.
    - ✧ If the above information cannot be provided prior to the use of AI systems and services, a system shall be established to respond to feedback from non-business users in accordance with the risks assumed based on the nature of AI and the manner of its use.

[Ref.]
- NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023)
- OECD, "Advancing accountability in AI" (February 2023).

[Description in this volume (reprinted)]
- When using AI systems and services

(U-7) i. Description to relevant stakeholders

✧ Provide information, including appropriate usage methods, in a plain and accessible manner to a reasonable extent depending on the nature of the relevant stakeholders ("7) Accountability")

- ✧ If it is planned to use data provided by relevant stakeholders, provide information to such stakeholders in advance on the means, format, etc. of data provision, taking into account the characteristics and uses of AI, points of contact with the recipient, privacy policy, etc. ("7) Accountability")
- ✧ If the output results of the AI are used as a reference for evaluation of a specific individual or group of individuals, the fact that the AI is being used shall be notified to the specific individual or group of individuals or groups being evaluated, the procedures to ensure the accuracy, fairness, transparency, etc. of the output results recommended in these Guidelines shall be observed, and accountability shall be ensured based on reasonable human judgment in consideration of automation bias (see "1. In addition, the AI shall be accountable upon request from the individuals or groups subject to the evaluation, based on reasonable human judgment, taking into account any automated biases ("1) Human-centeredness", "6) Transparency", and "7) Accountability").
- ✧ Depending on the nature of the AI system/service to be used, a point of contact should be established to respond to inquiries from relevant stakeholders and to receive explanations and requests in cooperation with the AI provider ((7) Accountability).

(U-7) ii. Use of documents provided and compliance with terms and conditions

- ✧ Appropriate storage and use of documentation about AI systems and services provided by AI providers ("7) Accountability")
- ✧ Comply with the terms of service established by the AI provider ("7) Accountability)

[Point]

AI users are expected to develop, publish, and provide notice of their policies on the use of AI so that non-business users are appropriately aware of the use of AI.

[Specific Methods]

- Disclosure of use policy on AI that includes the following
  - ➢ A statement that AI is being used (if specific functions and technologies can be identified, including the name and details of such functions and technologies)
  - ➢ Scope and Methods of AI Application
  - ➢ Basis of AI output
  - ➢ Risks associated with the use of AI
  - ➢ inquiry counter
  - ➢ Points to keep in mind when disclosing and notifying the Usage Policy
    - ✧ When AI is used in a way that its output directly affects non-business users or third parties, a policy on the use of AI should be prepared and disclosed so that non-business users and third parties can be appropriately aware of the use of AI, and an explanation should be provided if they inquire about it.
    - ✧ Proactively notify when there is a possibility of significant impact on the rights and interests of non-professional users or third parties. (It is considered that AI providers and AI users are required to publicize their use policies regarding AI when the output of the AI they use directly affects extra-business users or third parties. In other words, if AI is used only as an analytical tool for human thinking, or if AI is drafted but the final decision is substantially guaranteed to be made by a human being, publication of the usage policy on AI is not necessarily required. (However, it is expected that they will be voluntarily published.)
    - ✧ Notification or announcement is expected to be made not only before the start of use, but also when there is a change in the AI's operation or at the end of use (especially when there is a change in the assumed risk due to a change in the AI's operation, etc.).

- ✧ In the case of using AI to detect fraudulent activities or when there is a concern about the risk of abuse, determine whether, what, and how disclosure/notification is required before implementation.

[Ref.]
- NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" (January 2023)
- The White House, "Blueprint for an AI Bill of Rights (Making Automated Systems Work for The American People)" (October 2022).

B. Explanation of "Common Guiding Principles" in "Part 2"

This section describes specific methods that are not mentioned in "Part 5: Matters Concerning AI Users" of this volume, but are of particular importance to AI users among the "Common Guidelines" in "Part 2" of this volume.

[Contents of this volume (reprinted)] * Only the columns are excerpted.

1) anthropocentric

Each entity should ensure that the development, provision, and use of AI systems and services do not violate at least constitutionally guaranteed or internationally recognized human rights as a basis from which all matters to be addressed, including each of the items described below, are derived. It is also important to act in such a way that AI extends people's capabilities and enables the pursuit of diverse happiness (well-being) of diverse people.

- Related to (4) Ensuring Diversity and Inclusion
  [related description].
  - In addition to ensuring fairness, care will be taken to facilitate the use of AI by the socially vulnerable so that more people can enjoy the benefits of AI without creating so-called "information and technology weaklings".

  [Specific Methods]
  - A commitment to AI utilization that leaves no one behind.
    - ✧ Increase AI literacy
    - ✧ Secure and develop digital and AI human resources
    - ✧ Establishment of a safe and secure AI use environment

[Ref.]
- Ministry of Internal Affairs and Communications "White Paper on Information and Communications 2021" (July 2021)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(2) Safety

Each entity should ensure that the development, provision, and use of AI systems and services do not cause harm to the lives, bodies, or property of stakeholders. In addition, it is important to ensure that no harm is caused to the psyche and the environment.

- Consideration for human life, body, property, spirit, and environment
  [related description].
  - Examine measures to be taken in the event of a situation that compromises the safety of AI systems and services, and prepare for prompt implementation in the event of such a situation.

  [Specific Methods]
  - Organize incident countermeasures and consider measures to be taken in the event of an incident
    - Organizing Incident Response
      - Preparation of a communication system in the event of a hazardous situation
      - Organize methods for investigating causes and restoration work
      - Examine measures to prevent recurrence and organize response policies
      - Set up a method for sharing information about the incident
    - initial response measures
      - Recovery through rollback of AI systems, use of alternative systems, etc.
      - Deactivation of AI system (kill switch)
      - Disconnection of AI systems from the network
      - Confirmation of the nature of the harm
      - Reporting to relevant stakeholders
    - Use of insurance to facilitate compensation, indemnification, etc.
    - Establishment of a third-party organization to investigate, analyze, and make recommendations on the cause of the problem

[Ref.]
- Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry "Corporate Privacy Governance Guidebook in the DX Era ver1.3" (April 2023)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(3) Fairness

In the development, provision, and use of AI systems and services, it is important for each entity to strive to eliminate unfair and harmful prejudice and discrimination against specific individuals or groups on the basis of race, gender, nationality, age, political beliefs, religion, and other diverse backgrounds. It is also important for each entity to develop, provide, and use AI systems and services after assessing whether such unavoidable biases are acceptable from the perspective of respecting human rights and diverse cultures, while recognizing that some biases still cannot be avoided.

- Related to "(2) Intervention of human judgment"
  [related description].
  - To ensure that AI output results are not unbiased, consider using AI not only to make decisions on its own, but also to intervene with human judgment.

  [Specific Methods]
  - Decision-making regarding the need for human judgment intervention and ensuring its effectiveness
    - Judgment regarding the need for human judgment intervention (examples of judgment criteria are provided below)

- The nature of the rights and interests of AI users and off-business users affected by the output of AI, and the intentions of AI users and off-business users.
- Degree of reliability of the AI's output (superiority or inferiority to the reliability of human judgment)
- Timeframe required for human decision making
- Competencies expected of AI users and out-of-business users who make decisions
- The necessity of protection of the subject of the decision (e.g., whether to respond to individual applications by humans or to mass applications by AI systems/services, etc.)
- Uncertainty of statistical future predictions
- Necessity and degree of convincing reasons for decisions (judgments)
- Assumed degree of discrimination based on race, creed, and gender due to inclusion of social bias against minorities, etc. in the study data

➢ Ensuring the effectiveness of human judgment
  ✧ Clarify in advance the items on which humans should make judgments based on explanations obtained from the AI with explainability (when it is appropriate for humans to make final judgments on the AI's output).
  ✧ Acquire the necessary skills and knowledge to enable AI users and off-business users to appropriately judge AI outputs (when it is appropriate for humans to make the final judgment on AI outputs).
  ✧ Organize responses in advance to ensure the effectiveness of human decisions.

[Ref.]
- National Institute of Advanced Industrial Science and Technology, "Machine Learning Quality Management Guidelines, 4th Edition" (December 2023)

[Contents of this volume (reprinted)] * Only the columns are excerpted.

(8) Education and literacy

Each entity is expected to provide the necessary education to ensure that those involved in AI within the entity have the knowledge, literacy, and sense of ethics to have a correct understanding of AI and socially correct use of AI. Each entity is also expected to educate its stakeholders, taking into account the characteristics of AI such as complexity and misinformation, as well as the possibility of intentional misuse.

[related description].
- Take necessary steps to ensure that those involved in AI within each entity have a sufficient level of AI literacy in their involvement
- As the segregation of AI and human work is expected to change with the expanded use of generative AI, education, reskilling, etc. will be considered to enable new ways of working.

[Specific Methods]
- Items that should be included as literacy education and skills for using AI
  ➢ Knowledge of AI, mathematical and data science
  ➢ Understanding of the characteristics of AI and data, such as the presence of bias in the data and the potential for bias depending on how it is used
  ➢ An understanding of the challenges associated with AI and data in terms of fairness and privacy protection, as well as content related to security and the limitations of AI technology
  ➢ Understanding of the use of data and digital technologies, including AI, in a variety of operations
  ➢ Appropriate use of AI for productivity improvement by combining AI with skills such as "asking questions" and "formulating and testing hypotheses
- Education to improve resilience to environmental change, etc.

- Flexible mental rotation between "vertical thinking" and "horizontal thinking
- Strengthen modeling skills required for organizational assessment
- Extending learning horizons in skill areas that are not well served by agile thinking
- Improved valuation skills to analyze uncertainties that are difficult to predict with past experience and expertise
- Conversion of "key points" of organizational evaluation (to "instantaneous + sustained" mobility)
- How to evaluate management resynchronization (Configuration, Architecture, Synthesis, Dissemination), which is becoming increasingly complex and sophisticated as forms of AI governance become increasingly hybrid, centralized and decentralized

[Ref.]

- Ministry of Economy, Trade and Industry, Information-technology Promotion Agency, Japan "Digital Skill Standards ver.1.1" (August 2023)
- Cabinet Office, "Principles for a Human-Centered AI Society" (March 2019)

Appendix 6. Main considerations when referring to the "Contractual Guidelines for the Use of AI and Data".

As explained in "Part 2" of this volume, the development, provision, and use of AI involve multiple entities in each situation. Therefore, it is expected that the rights and obligations of the parties involved in each transaction related to the development, provision, and use of AI should be stipulated as clearly as possible in the contract, and that guidelines for resolving disputes should they arise, in order to facilitate each transaction and prevent needless disputes that may arise.

The first edition of the Contract Guidelines for the Use of AI and Data was developed and published in June 2018 at[78] (Version 1.1 with partially updated content was published in December 2019. Hereinafter referred to as the "Contract Guidelines")), against the backdrop of the issues at the time, lays out basic ideas on contracts for the development and use of AI-based software and contracts for the provision/use of data, as well as matters that should be understood in advance as preconditions for these contracts.

The contract guidelines were developed in the midst of the trend toward the future development and practical application of AI, and the following issues were identified as those that should be resolved through this process under the objective of the guidelines to encourage the development and use of AI.

- Lack of accumulated practical experience in contracts for the provision/use of AI and data
- Gaps in recognition and understanding between the parties regarding the technical characteristics of AI and the value of data and AI development know-how.

At the time the contract guidelines were formulated, awareness of the issue was also placed on removing obstacles to such transactions, with the aim of facilitating transactions between those who develop AI-based software and those who use the fruits of such development, thereby encouraging the development and practical application of AI.

Five years have already passed since the formulation and publication of the first edition of the Contract Guidelines, but in the intervening period, the situation regarding the development and use of AI has progressed remarkably, with new technologies and methods of use being created daily, and many technologies related to AI have entered a phase where they are on their way to widespread use in society. Due to these developments, it is important to keep in mind that there are some contents in the contract guidelines that are still beneficial to refer to, and others that should be considered in light of changes in the situation after their publication.

As an example, among the descriptions in the Contract Guidelines, the following contents, which are mainly referred to in the AI Section 2 (Description of AI Technology) and 3 (Basic Concepts) and the Data Section 3 (Legal Basics for Considering Data Contracts), may be generally useful to refer to as before. The following information, which are referred to in the Data Chapter 3 (Commentary on AI Technology) and the Data Chapter 3 (Legal Basics for Considering Data Contracts), may be of general interest.

[78] Ministry of Economy, Trade and Industry, "Contract Guidelines for the Use of AI and Data, Version 1.1" (accessed December 2019), https://warp.da.ndl.go.jp/info:ndljp/pid/12685722/www.meti.go.jp/press/2019/12/ 20191209001/20191209001-1.pdf

(AI ed.)
- Characteristics of AI and AI-based software development
- Arrangement of Intellectual Property Rights, etc.
- Basic perspectives on attribution of rights and establishment of conditions of use
- Basic Perspectives on Distribution of Responsibility

(Data ed.)
- Legal nature of data and protection measures for data

In addition, it is still considered beneficial to refer to the explanations of the various contract models dealt with in the Contract Guidelines; however, transactions related to the development, provision, and use of AI have become more diverse than when the Contract Guidelines were established, and the differences between the various contract models and actual transactions should be The differences between the various contract models and actual transactions need to be carefully considered.

On the other hand, it is important to note the following points, in particular, as matters that should be considered in light of changes in circumstances after the publication of the Contract Guidelines.

(1) Concept of AI development and utilization

Based on the premise of the dichotomy between those who develop AI (vendors) and those who use AI (users), the Contract Guidelines provide two model contracts, a development contract and a use contract, for each type of transaction: (1) a transaction in which a user commissions a vendor to develop AI and (2) a transaction in which a user is allowed to use AI developed by a vendor, and The paper provides explanations of the two model contracts, development contract and utilization contract, for each type of transaction.

It is thought that some transactions related to the development and use of AI may still exist today, with some of these arrangements remaining unchanged as they are. However, in the recent trend from development to commercialization and from diffusion to application of AI, society's interest has shifted from what kind of technology to develop to how to use the technology, and transactions that do not fit into the types of transactions outlined in the contract guidelines are becoming increasingly important.

(Example transaction)
- Transactions related to the development of software incorporating AI
- Transactions related to AI maintenance and operations
- Transactions related to the optimization of AI for specific purposes
- Transactions centered on consulting on the use of AI and data.

The model contracts in the Contract Guidelines cannot be used as they are in relation to such transactions, and must be examined in accordance with the actual conditions of each transaction. For example, when the person who develops AI and the person who maintains and operates the development results are different, there may be a trade-off between protecting the know-how of AI development and maintaining and operating the development results. In dealing with such a situation, it is necessary to find a solution that is in line with the actual situation, while referring to the description in the contract guidelines as far as it is relevant.

(2) Development, provision, and use of AI and allocation of responsibility

In the contract guideline, in both types of transactions, (1) where the user entrusts the vendor with the development of AI and (2) where the vendor allows the user to use the AI developed by the vendor, the description is devoted to organizing the legal relationship between the vendor and the user, especially the attribution and use relationship of results, with a simple interest model that can be adjusted between the vendor and the user. The description is devoted to organizing the ideas concerning the burden of risk of damage to the parties and infringement of intellectual property rights and other rights of third parties.

Recently, however, the value chain of AI has become more diverse and complex compared to the time when the contract guideline was established, as the following various businesses are now involved in the development, provision, and use of AI. In addition, the increasing prevalence of AI has forced us to be aware of the existence of non-business users, and as a result, problems have arisen that cannot be adequately captured by simply focusing on the bilateral relationship between the vendor and the user.

(Example of a business)
- Entity that develops AI (AI developer in these guidelines)
- Entities that develop software incorporating the developed AI (AI developers in these guidelines)
- The entity that provides the software externally (in this guideline, the AI provider)
- Entities that provide services using the provided software to external parties (AI providers in this guideline)
- The entity that uses the service (AI user in this guideline)

One such issue is how responsibility should be distributed along the AI value chain. For example, in cases where services using AI-embedded software are provided to non-business users, the question may arise as to who should assume the risk of damage caused by the AI to the non-business users. The content and degree of such risks are greatly affected not only by the quality of the AI but also by the way the software is provided and the way the service is provided and used. Without looking beyond the bilateral relationship between vendor and user and into the role of each party in the value chain, it can be difficult to establish reasonable boundaries for the scope of their responsibilities.

Related to this boundary rationality is the issue of assumption of risk by a party that has no direct control over the risk. Taking the above case as an example, if the AI developer assumes all responsibility for any damage caused by the AI, regardless of how it is provided or used, the AI developer will also assume risks that it cannot directly control. While such a situation is likely to occur between parties with disparities in bargaining power, there are cases in which the scope of risks that should be controlled by the AI developer should be broadly considered if the AI has a large influence.

The Contract Guidelines focus on bilateral transactions between a vendor and a user, and the distribution of responsibility based on such a diversified or complicated value chain should be considered according to individual situations. In this regard, the practical examples in Appendix 2, 3. System Design (Construction of AI Management System), may be helpful.

(3) Development, provision, use and accountability of AI

As AI becomes more prevalent and applied, the risks associated with the development, provision, and use of AI will increase, and the number of cases in which such risks become apparent will also increase in the future.

Among such risks is the risk that an accident may occur in relation to software incorporating AI or services using AI, resulting in damage to parties involved in the development, provision, or use of AI, or to third parties. The parties involved in the development, provision, and use of

AI may be required to provide a reasonable explanation of their involvement in the process. The parties involved in the development, provision, and use of AI may be required to provide reasonable explanations for their involvement in each of these processes.

Liability for such an explanation is one that may arise for the party that bears primary responsibility for the accident, no matter what contract has been entered into between all parties to the development, provision, and use of the AI. The contract can only provide for the allocation of liability as far as the parties to the contract are concerned. All parties linked to the AI value chain could stand to be held to certain account in the event that liability is sought by parties other than the parties to the contract.

What matters in relation to the reasonableness of explanations is the objective basis for such explanations in addition to the content of the explanations, and it is expected that such basis should be organized before and after the conclusion of a contract for the development, provision, and use of AI. Although not mentioned in the contract guideline, it would be beneficial to refer to the practical examples regarding 3. system design (establishment of AI management system) in Appendix 2 and consider how to deal with such issues after the conclusion of the contract.

The development and diffusion of technologies related to AI has been remarkable, and new technologies and methods of use are being created every day, and the points to be considered in contracts are changing accordingly. In considering contracts, it is important to consider how contracts should be made and the risks involved in each transaction of development, provision, and use of AI, and refer to the "Guidelines for Contracts on AI and Data Use" in light of the above-mentioned points to be considered.